LD-DETR：用於視頻時刻檢索和精彩片段檢測的循環解碼器檢測 Transformer

趙鵬程，賀之纖，張富為，林淑金*，周凡

中山大學，中國廣州

qingchen239@gmail.com, hezhx29@mail2.sysu.edu.cn,
zhangfw5@mail2.sysu.edu.cn, linshjin@mail.sysu.edu.cn, isszf@mail.sysu.edu.cn

摘要

視頻時刻檢索與精彩片段檢測旨在根據文本查詢找到視頻中的對應內容。現有模型通常首先使用對比學習方法來對齊視頻和文本特徵，然後融合並提取多模態信息，最後使用 Transformer 解碼器解碼多模態信息。然而，現有方法面臨幾個問題：(1) 數據集中不同樣本之間重疊的語義信息阻礙了模型的多模態對齊性能；(2) 現有模型無法有效提取視頻的局部特徵；(3) 現有模型使用的 Transformer 解碼器無法充分解碼多模態特徵. 針對上述問題，我們提出了用於視頻時刻檢索和精彩片段檢測任務的 LD-DETR 模型。具體而言，我們首先將相似度矩陣提取到單位矩陣以減輕重疊語義信息的影響。然後，我們設計了一種方法，使卷積層能夠更有效地提取多模態局部特徵。最後，我們將 Transformer 解碼器的輸出反饋到其自身中，以充分解碼多模態信息。我們在四個公共數據集上對 LD-DETR 進行了評估，並進行了廣泛的實驗，以證明我們的方法的優越性和有效性。我們的模型在 QVHighlight、Charades-STA 和 TACoS 數據集上的表現優於最先進的模型 (State-Of-The-Art)。我們的代碼可在以下網址獲取：https://github.com/qingchen239/ld-detr

1 介紹

視頻時刻檢索旨在識別視頻中與給定文本查詢相對應的特定時刻 (Liu et al. 2015; Anne Hendricks et al. 2017; Gao et al. 2017; Liu et al. 2018; Escorcia et al. 2019b)。精彩片段檢測評估不同時間片段與文本的相關程度 (Yao, Mei, and Rui 2016; Zhang et al. 2016; Gygli, Song, and Cao 2016; Yu et al. 2018; Xiong et al. 2019)。隨著數字設備和平台的發展，用戶對於視頻內容的需求大幅增加，快速準確地找到視頻中有趣的片段成爲重要需求，因此視頻時刻檢索與精彩片段檢測的研究受到廣泛關注。

現有的視頻時刻檢索和精彩片段檢測模型往往使用對比學習來對齊視頻和文本特徵 (Moon et al. 2023b; Sun et al. 2024; Moon et al. 2023a; Liu et al. 2024b; Xiao et al. 2024)，使用注意力機制來融合和提取多模態信息 (Liu et al. 2022b; Moon et al. 2023b; Sun et al. 2024)，並使用 Transformer 解碼器和一個作爲查詢 (*query*) 的零矩陣來解碼融合的多模態信息 (Zheng et al. 2020; Lei, Berg, and Bansal 2021).

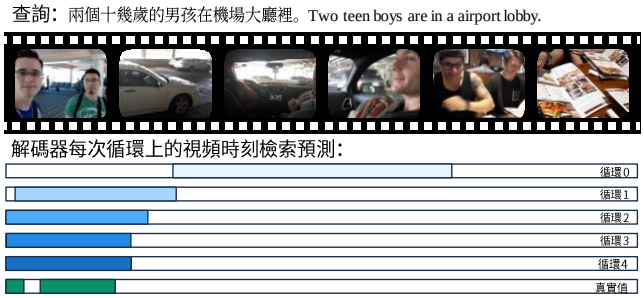

圖 1: 循環解碼器使視頻時刻檢索更加準確。我們將每次循環時循環解碼器輸出對應的視頻時刻檢索結果可視化。隨著循環次數的增加，結果越來越接近眞實值。此實驗以 M-DETR (Lei, Berg, and Bansal 2021) 爲基準模型。

然而，現有方法面臨幾個問題：(1) 在對比學習中，方法通常將來自不同樣本的特徵視爲負樣本 (Sun et al. 2024; Moon et al. 2023a)，但相同的語義信息不可避免地出現在不同的樣本中 (例如，「人吃飯」和「人喝水」都具有相同的信息「人」) (Jung et al. 2023)。將它們簡單地視爲完全的負樣本會阻礙多模態對齊的性能。(2) 相應的內容通常是視頻的一小部分，具有很強的局部相關性，但當前的模型忽略了提取視頻的局部特徵。一種直觀的方法是使用卷積層 (LeCun et al. 1998; Krizhevsky, Sutskever, and Hinton 2012) 來提取局部特徵 (Xiao et al. 2024; 黎金宇 2024)，但簡單地堆疊卷積層並不能提高現有模型的性能。(3) 研究表明，Transformer 解碼器不足以處理檢索 (Yang et al. 2024; Liu et al. 2022a; Zhu et al. 2020; Gao et al. 2021b; Meng et al. 2021; Wang et al. 2022a; Yao et al. 2021)。更大的解碼器可能會提高模型的檢索能力，但也存在過擬合的風險。

*通訊作者。

爲了解決這些問題，我們提出了循環解碼器檢測 Transformer 模型（Loop Decoder DEtection TRansformer，LD-DETR）：（1）與其他以單位矩陣作爲目標的對比學習方法不同，我們將一個表示樣本間相關性的矩陣提取到單位矩陣中，以減輕語義信息重疊的影響。（2）卷積層的感受野較小，這一特性使得網絡能夠捕獲局部信息。我們設計了一種方法，使得堆疊的卷積層能夠更高效地提取多模態局部特徵。（3）研究表明，當 Transformer 解碼器的查詢（*query*）攜帶目標信息時，解碼器能夠更好地解碼輸入信息 (Liu et al. 2022a)，而 Transformer 解碼器的輸出也恰恰攜帶目標信息。受此啓發，我們將 Transformer 解碼器的輸出作爲查詢反饋到自身，如圖 1 所示，以增強其充分解碼多模態融合信息的能力，同時又不增加過擬合的風險。

我們在四個公共數據集上對 LD-DETR 進行了評估，並進行了大量實驗以證明我們方法的優越性和有效性。

我們工作的主要貢獻總結如下：

- 我們提出了一種即插即用的方法提取對齊（Distill Align），該方法在對齊多模態特徵時考慮訓練樣本之間重疊語義信息的影響，以提高模型的性能。
- 我們引入了卷積融合器（Convolutional Fuser）來更好地提取視頻中的局部特徵並取得了優異的效果。
- 我們提出了一種即插即用的方法，循環解碼器（Loop Decoder），它提高了解碼器充分解碼多模態融合信息的能力，而不會導致過擬合。
- 基於以上方法，我們設計了用於視頻時刻檢索和精彩片段檢測任務的模型 LD-DETR，並在多個數據集上驗證了其先進性和有效性。我們的模型在 QVHighlight、Charades-STA 和 TACoS 數據集上的表現優於最先進的模型。

2 相關工作

自 QVHighlight 數據集 (Lei, Berg, and Bansal 2021) 提出以來，視頻時刻檢索和亮點檢測任務被聯合研究，並提出了許多基於檢測 Transformer（DEtection TRansformer，DETR）(Zheng et al. 2020) 的模型。這些方法主要從對齊多模態特徵、融合提取多模態特徵、解碼多模態信息三個角度對模型進行改進。

2.1 對比學習和對齊多模態特徵

對比學習是一種通過比較和對比不同樣本來提高其區分特徵和識別模式的性能的機器學習方法。CMC (Tian, Krishnan, and Isola 2020) 利用對比學習將同一幅圖像的不同視角映射到相似的語義空間中，證明了對齊多模態信息的可行性。MoCo (He et al. 2020) 通過動量更新編碼器獲取負樣本的特徵並將其存儲在隊列中，以增加參與對比學習的樣本數量。CLIP (Radford et al. 2021) 爲每個圖像和文本提取全局特徵，通過確保它們的相關矩陣近似於單位矩陣來對齊它們。

現有的視頻時刻檢索和精彩片段檢測模型一般採用類似 CLIP 的方法對視頻和文本特徵進行對齊。TR-DETR (Sun et al. 2024) 將單模態編碼器編碼後的視頻和文本特徵的平均值作爲樣本的全局特徵，再採用類似 CLIP 的方法對多模態特徵進行對齊，最後再進行混合。CG-DETR (Moon et al. 2023a) 採用兩個不同的編碼器分別提取正樣本和負樣本的視頻和文本的全局特徵，然後通過類似 CLIP 的方法分別進行對齊。在這些類似 CLIP 的方法中，對比學習所涉及的特徵數量受到批次大小（batch size）的限制。

BM-DETR (Jung et al. 2023) 提出了弱對齊問題，數據集中不同樣本之間重疊的語義信息降低了模型性能，並在多模態特徵融合層面解決該問題。我們認爲該問題也可能降低多模態對齊，我們將在多模態對齊層面解決該問題。

2.2 融合和提取多模態特徵

Moment-DETR (Lei, Berg, and Bansal 2021) 只是將視頻和文本特徵拼接起來，然後送入 Transformer 編碼器。UMT (Liu et al. 2022b) 和 QD-DETR (Moon et al. 2023b) 提出了交叉注意力 Transformer 編碼器，利用文本特徵對視頻特徵進行編碼，以去除視頻特徵中與查詢文本無關的信息。CG-DETR (Moon et al. 2023a) 更進一步，在交叉注意力 Transformer 編碼器的基礎上將噪聲拼接到文本特徵上，以更好地去除無關的信息。TR-DETR (Sun et al. 2024) 引入了視覺特徵細化來過濾掉無關的視頻信息。現有的這些方法只注重去除視頻特徵中與文本無關的特徵，而忽略了視頻本身的時間結構，將提取視頻特徵的任務留給了 Transformer 編碼器 (Vaswani et al. 2017)。但 Transformer 編碼器的全局注意力計算 (Bahdanau, Cho, and Bengio 2014) 稀釋了局部細節的權重，導致模型忽略了視頻局部特徵的提取。UVCOM (Xiao et al. 2024) 提出了綜合集成模塊方法，其中有一個卷積層來提取局部多模態特徵。CDIM (黎金宇 2024) 還提出了一種跨模態卷積交互方法，該方法堆疊擴張卷積層以增強模型的感知能力。

2.3 解碼多模態信息

現有的模型大多基於一個應用於物體檢測的模型 DETR (Zheng et al. 2020)，研究 (Yang et al. 2024; Liu et al. 2022a; Zhu et al. 2020; Gao et al. 2021b; Meng et al. 2021; Wang et al. 2022a; Yao et al. 2021) 表明 Transformer 解碼器對融合多模態信息的處理不夠充分，並提出了適合物體檢測任務的解決方案。MomentDiff (Li et al. 2024) 注意到訓練樣本分佈不均導致模型泛化能力不足，在解碼器中引入擴散模型的方法解決這個問題。

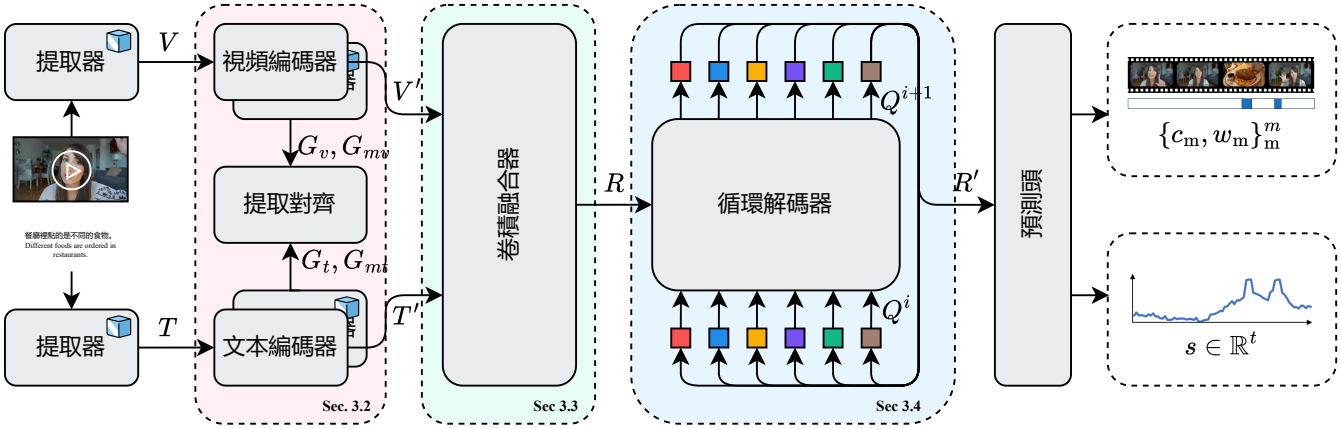

圖 2: 我們的模型 LD-DETR 的總體框架。對於標有冰塊的方法，其參數在訓練過程中不受梯度下降的影響。

UVCOM (Xiao et al. 2024) 利用文本和視頻特徵的交叉注意作爲 Transformer 解碼器的查詢。TaskWeave (Yang et al. 2024) 使用類似 DAB-DETR (Liu et al. 2022a) 的解碼器解碼視頻時刻檢索特徵，使用另一個解碼器解碼精彩片段檢測特徵。

3 模型和方法

3.1 LD-DETR 概述

給定一個包含 t 個片段（clip）的視頻和一個包含 n 個詞（token）的文本查詢，視頻時刻檢索和精彩片段檢測的目標是找到與視頻中的文本查詢相關的所有時刻 $\{c_m, w_m\}_m^m$，其中 c_m 和 w_m 表示第 m 個時刻的中心時間和持續時間長度，m 是預測時刻的總數，並預測片段級別的所有時刻的顯著性得分 $s \in \mathbb{R}^t$。

如圖 2 所示，LD-DETR 可分爲五個部分：單模態編碼器（Unimodal Encoder）、提取對齊（Distill Align）、卷積融合器（Convolutional Fuser）、循環解碼器（Loop Decoder）和預測頭（Prediction Heads）。

輸入的視頻和文本首先被輸入到預先訓練的特徵提取器中，以提取視頻和文本特徵 $V \in \mathbb{R}^{b\times n\times d_v}, T \in \mathbb{R}^{b\times n\times d_t}$，其中 b 是批量大小，d_v 和 d_t 分別是提取的視頻片段和文本標記特徵的維度。然後將視頻和文本特徵輸入到兩個單模態編碼器 $\mathrm{UE}_v(\cdot), \mathrm{UE}_t(\cdot)$ 中，以映射到潛在空間 $V' \in \mathbb{R}^{b\times t\times d}, T' \in \mathbb{R}^{b\times n\times d}$，其中 d 是模型的隱藏維度。我們使用兩個單模態編碼器和兩個動量單模態編碼器 $\mathrm{UE}_{vm}(\cdot), \mathrm{UE}_{tm}(\cdot)$ 獲取全局特徵。將單模態編碼器獲取的兩個全局特徵和兩個動量全局特徵 $G_v, G_{mv}, G_t, G_{mt} \in \mathbb{R}^{b\times d}$ 輸入到提取對齊方法中，保證特徵映射到同一個空間。然後將映射後的特徵 V', T' 輸入到卷積融合器中，得到多模態特徵 $R \in \mathbb{R}^{b\times t\times d}$。然後將多模態特徵 R 和零矩陣 $O \in \mathbb{R}^{b\times q\times d}$ 一起送入循環解碼器，得到解碼特徵 $R' \in \mathbb{R}^{b\times q\times d}$，其中 q 爲超參數，表示參考點的數量。最後將解碼特徵送入預測頭，得到預測時間 $\{c_m, w_m\}_m^m$ 和顯著性得分 $s \in \mathbb{R}^t$。

3.2 單模態編碼器（Unimodal Encoder）和提取對齊（Distill Align）

我們使用一個兩層的多層感知器 (Rumelhart, Hinton, and Williams 1986) 作爲單模態編碼器 $\mathrm{UE}(\cdot)$ 將提取的特徵 $X \in \mathbb{R}^{b\times x\times d_x}$ 映射到潛在空間 $X' \in \mathbb{R}^{b\times x\times d}$，其中 $X \in \{V, T\}$ 同時 $x \in \{t, n\}$。我們利用樣本在潛在空間 X' 中在片段（clip）或詞（token）維度上的特徵的平均值來獲得每個樣本的全局特徵 $G \in \mathbb{R}^{b\times d}$。

我們利用兩個可學習的單模態編碼器 $\mathrm{UE}(\cdot) \in \{\mathrm{UE}_v(\cdot), \mathrm{UE}_t(\cdot)\}$ 和兩個動量單模態編碼器 $\mathrm{UE}_m(\cdot) \in \{\mathrm{UE}_{mv}(\cdot), \mathrm{UE}_{mt}(\cdot)\}$ 將提取的特徵 $V \in \mathbb{R}^{b\times x\times d_v}, T \in \mathbb{R}^{b\times x\times d_t}$ 映射到潛在空間，分別得到四個全局特徵 G_v, G_{mv}, G_t, G_{mt}。動量單模態編碼器由相應的單模態編碼器更新：

$$\mathrm{UE}^0_{m\theta} = \mathrm{UE}^0_{\theta}, \tag{1}$$

$$\mathrm{UE}^i_{m\theta} = m\mathrm{UE}^{i-1}_{m\theta} + (1-m)\mathrm{UE}^i_{\theta}, \quad \text{when } i > 0, \tag{2}$$

其中 $m \in [0, 1)$ 爲動量（momentum）係數，X_θ 表示方法 X 中的所有參數。得到映射特徵 $V', V'_m, T', T'_m \in \mathbb{R}^{b\times x\times d}$ 後，我們計算每個樣本的全局特徵 $G_v, G_{mv}, G_t, G_{mt} \in \mathbb{R}^{b\times d}$。我們將動量全局特徵推送到動量全局特徵隊列 $Q_v, Q_t \in \mathbb{R}^{l\times d}$，其中 l 表示隊列長度。

圖 3 展示了提取對齊的結構。此時，我們有視頻全局特徵 $G_v \in \mathbb{R}^{b\times d}$、視頻和文本動量全局特徵 $G_{mv}, G_{mt} \in \mathbb{R}^{b\times d}$，以及文本動量全局特徵隊列 $Q_t \in \mathbb{R}^{l\times d}$。通過計算餘弦相似度 $s(\cdot,\cdot)$，我們得到了這些全局特徵中的視頻到文本相似度矩陣 $S_{v2t}, S_{v2tm} \in \mathbb{R}^{b\times l}$：

$$S_{v2t} = s(G_v, Q_t), \tag{3}$$

$$S_{v2tm} = s(G_{mv}, Q_t). \tag{4}$$

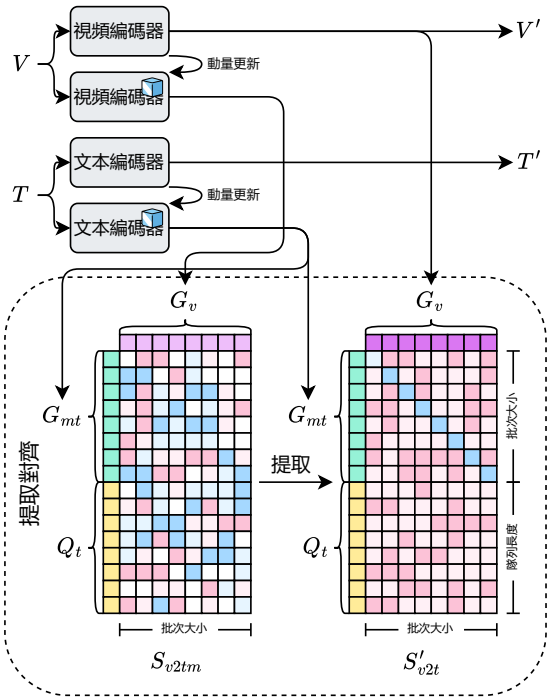

圖 3: 提取對齊的結構。

爲了減輕語義信息重疊的影響，我們將相似度矩陣 $S_{v2tm} \in \mathbb{R}^{b\times l}$ 提取到單位矩陣 I，作爲多模態比對的目標 $S'_{v2t} \in \mathbb{R}^{b\times l}$：

$$S'_{v2t} = \alpha S_{v2tm} + (1-\alpha)I, \tag{5}$$

其中 $\alpha \in [0,1)$ 是提取係數（distillation coefficient）。最後，視頻到文本對齊損失 $\mathscr{L}_{v2t}$ 是

$$\mathscr{L}_{v2t} = \mathrm{CE}(S_{v2t}, S'_{v2t}), \tag{6}$$

其中 $\mathrm{CE}(\cdot,\cdot)$ 是交叉熵損失（cross entropy loss）。

類似地，我們計算文本到視頻的對齊損失 $\mathscr{L}_{t2v}$。最終的多模態對齊損失爲

$$\mathscr{L}_{align} = (\mathscr{L}_{v2t} + \mathscr{L}_{t2v})/2. \tag{7}$$

3.3 卷積融合器 (Convolutional Fuser)

圖 4 顯示了卷積融合器的結構。首先，我們將映射到潛在空間的視頻特徵 $V' \in \mathbb{R}^{b\times t\times d}$ 和文本特徵 $T' \in \mathbb{R}^{b\times n\times d}$ 輸入到 V2T 提取器（V2T Extractor）(Sun et al. 2024)、T2V 編碼器（T2V Encoder）(Moon et al. 2023b) 和 Transformer 編碼器 (Vaswani et al. 2017)，以獲得與文本無關的視頻特徵 $V'' \in \mathbb{R}^{b\times t\times d}$。隨後，將與文本無關的視頻特徵 $V'' \in \mathbb{R}^{b\times t\times d}$ 輸入到卷積塊（Convolutional Blocks）中。經過這個殘差網絡後，我們得到局部增強的多模態融合特徵 $V''' \in \mathbb{R}^{b\times t\times d}$。最後，將局部增

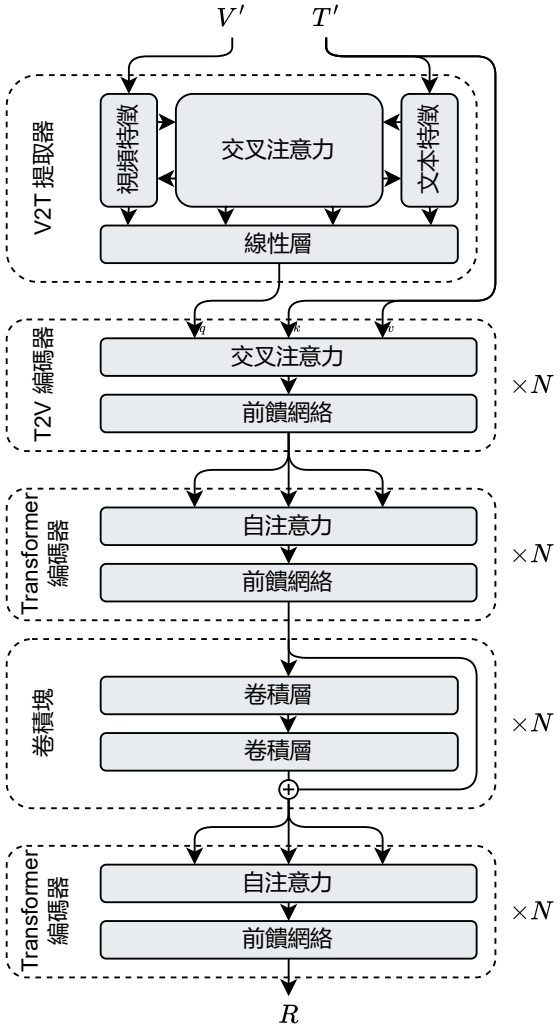

圖 4: 卷積融合器的結構。

強的多模態融合特徵 $V''' \in \mathbb{R}^{b\times t\times d}$ 輸入到另一個 Transformer 編碼器 (Vaswani et al. 2017) 中，得到多模態特徵 $R \in \mathbb{R}^{b\times t\times d}$。

V2T 提取器 (V2T Extractor) 首先，我們利用映射到潛在空間的視頻特徵 $V' \in \mathbb{R}^{b\times t\times d}$ 和文本特徵 $T' \in \mathbb{R}^{b\times n\times d}$ 得到視頻特徵與文本特徵之間的相關矩陣 $A \in \mathbb{R}^{b\times t\times n}$：

$$A_1 = \mathrm{Linear}(V'), \tag{8}$$
$$A_2 = \mathrm{Linear}(T'), \tag{9}$$
$$A_3 = \mathrm{Linear}(V')T'^T, \tag{10}$$
$$A = A_1 + A_2^T + A_3, \tag{11}$$

當矩陣相加時，較短的維度會自行擴展爲與其他矩陣相同的長度。然後我們在文本和視頻維度上對相關矩陣 A 執行 softmax，得到另外兩個相關矩陣 $A_r, A_c \in \mathbb{R}^{b\times t\times n}$。然後我

們得到與文本無關的視頻特徵 $V_v' \in \mathbb{R}^{b\times t\times d}$：

$$T_v = A_r T', \tag{12}$$
$$V_t = A_t A_c^T V', \tag{13}$$
$$V_{cat} = \left[V' || T_v || V' \circ T_v || V' \circ V_t\right], \tag{14}$$
$$V_{cat}' = \text{Linear}(V_{cat}), \tag{15}$$
$$B = \text{Softmax}(\text{Linear}(T')), \tag{16}$$
$$T_p = T'^T B, \tag{17}$$
$$V_{cat}'' = \left[V_{cat}' || T_p\right], \tag{18}$$
$$V_v' = \text{Linear}(V_{cat}''), \tag{19}$$

其中 $[\cdot||\cdot]$ 表示連接，$\circ$ 表示哈達瑪積。

V2T 編碼器 (T2V Encoder) 然後，我們利用與文本無關的視頻特徵 V_v' 來獲得文本引導的與文本無關的視頻特徵 $V'' \in \mathbb{R}^{b\times t\times d}$：

$$Q_v = \text{Linear}(V_v'), \tag{20}$$
$$K_t = \text{Linear}(T'), \tag{21}$$
$$V_t = \text{Linear}(T'), \tag{22}$$
$$\text{Attention}(Q_v, K_t, V_t) = \text{Softmax}\left(\frac{Q_v K_t^T}{d}\right) V_t, \tag{23}$$
$$V'' = \text{FFN}(\text{Attention}(Q_v, K_t, V_t)). \tag{24}$$

Transformer 編碼器 這裡的 Transformer 編碼器和其他論文裡的沒什麼區別 (Vaswani et al. 2017)。需要注意的是，這裡的兩個 Transformer 編碼器並不共享參數。

卷積塊 (Convolutional Blocks) 隨後，將與文本無關的視頻特徵 $V'' \in \mathbb{R}^{b\times t\times d}$ 輸入卷積塊。根據之前在圖像識別中的工作，我們使用類似於 ResNet (He et al. 2016) 的殘差塊 $\text{RB}(\cdot)$：

$$X_i = \text{RB}_i(X_{i-1}) = \sigma(X_{i-1} + \mathscr{F}(X_{i-1})), \tag{25}$$
$$\mathscr{F}(X_{i-1}) = \text{BN}(\text{Conv}(\sigma(\text{BN}(\text{Conv}(X_{i-1}))))), \tag{26}$$

其中 $\text{Conv}(\cdot)$ 表示卷積層，$\text{BN}(\cdot)$ 表示批量歸一化，$\sigma(\cdot)$ 表示激活函數。我們堆疊 N 個殘差塊來提取視頻特徵中的局部信息：

$$X_0 = V'', \tag{27}$$
$$X_i = \text{RB}_i(X_{i-1}), \qquad \text{when } i > 0, \tag{25}$$
$$V''' = X_N. \tag{28}$$

經過這個殘差網絡後，我們得到局部增強的多模態融合特徵 $V''' \in \mathbb{R}^{b\times t\times d}$。

3.4 循環解碼器 (Loop Decoder)

研究表明，當 Transformer 解碼器的查詢 (*query*) 攜帶目標信息時，解碼器能夠更好地解碼輸入信息 (Liu et al. 2022a)，而 Transformer 解碼器的輸出也恰恰攜帶目標信息。受此啓發，我們將 Transformer 解碼器的輸出作爲查詢反饋到自身，以增強其解碼多模態融合信息的能力。

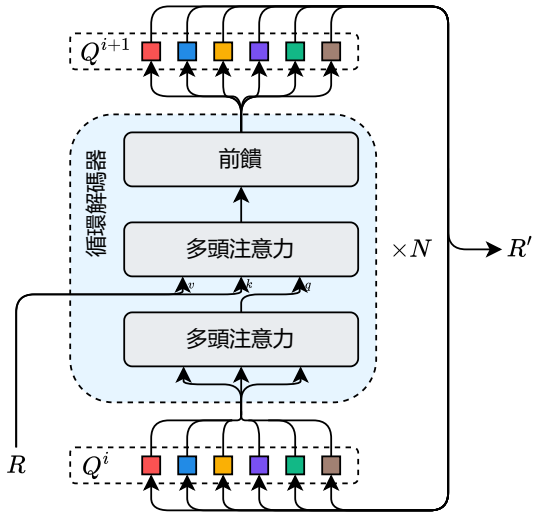

圖 5: 循環解碼器的結構。

圖 3 展示了循環解碼器的結構。我們將一個零矩陣 $O \in \mathbb{R}^{b\times q\times d}$ 和多模態特徵 $R \in \mathbb{R}^{b\times t\times d}$ 送入 Transformer 解碼器 $\text{TD}(\cdot,\cdot)$ (Vaswani et al. 2017)。經過 Transformer 解碼器的 N 次循環後，得到解碼特徵 $R' \in \mathbb{R}^{b\times q\times d}$：

$$Q^0 = O, \tag{29}$$
$$Q^i = \text{TD}(Q^{i-1}, R), \qquad \text{when } i > 0, \tag{30}$$
$$R' = Q^N, \tag{31}$$

其中 q 爲超參數，代表參考點的數量。與更大的解碼器相比，循環解碼器沒有引入新的參數，不會引起過擬合。

3.5 預測頭 (Prediction Heads)

我們使用類似於 QD-DETR (Moon et al. 2023b) 和 TR-DETR (Sun et al. 2024) 的預測頭。

視頻時刻檢索預測頭 (Video Moment Retrieval Prediction Head) 我們將解碼後的特徵 $R' \in \mathbb{R}^{b\times q\times d}$ 輸入多層感知器，得到 q 個預測時間 $\{c_\text{m}, w_\text{m}\}_\text{m}^q$：

$$\{s, e\} = \text{MLP}(R'), \tag{32}$$

其中 s 和 e 分別表示一個預測時間的開始和結束。同時利用解碼後的特徵 R 和多模態特徵 $R' \in \mathbb{R}^{b\times t\times d}$ 得到每個預測時間 $p \in \mathbb{R}^q$ 的置信度：

$$p = \text{MLP}(R') + \text{Sigmoid}^{-1}(\text{MLP}(R')), \tag{33}$$

其中 $\text{Sigmoid}^{-1}(\cdot)$ 表示反 S 型函數。這樣，我們就得到了視頻中與文本查詢相關的所有時間 $\{c_\text{m}, w_\text{m}\}_\text{m}^m$。

精彩片段檢測預測頭 (Highlight Detection Prediction Head) 得到所有時間之後，我們取出時間中所有片段（clip）對應的解碼特徵 $R'' \in \mathbb{R}^{b\times t'\times d}$，其中 t' 表示視頻時刻檢索預測中的片段總數，將它們輸入到門控循環單元 (Chung et al. 2014)，並使用隱藏狀態作爲視頻 $G'_v \in \mathbb{R}^{b\times d}$ 的新全局特徵：

$$o = \text{GRU}(R''), \tag{34}$$

$$G'_v = \text{GRU}_\theta, \tag{35}$$

其中 o 表示我們不關心的門控循環單元的輸出，GRU_θ 表示門控循環單元的隱藏狀態。然後，我們計算新的全局特徵 G'_v 與每個剪輯的特徵 V' 之間的相似度：

$$S = G'_v V'^T \tag{36}$$

最後得到顯著性分數 $s \in \mathbb{R}^t$：

$$M = \text{Linear}(R' \circ S + R'), \tag{37}$$

$$s = \text{sum}(M)/d, \tag{38}$$

其中 $\circ$ 表示哈達瑪積，$\text{sum}(\cdot)$ 表示對矩陣元素進行列求和，d 是模型的隱藏維度。

3.6 目標損失

LD-DETR 的目標損失函數 $\mathscr{L}_{total}$ 爲

$$\mathscr{L}_{total} = \mathscr{L}_{mr} + \mathscr{L}_{hd} + \lambda_{aligh}\mathscr{L}_{align}, \tag{39}$$

其中

$$\mathscr{L}_{mr} = \mathscr{L}_{mom} + \lambda_{CE}\text{CE}(\hat{y}, y), \tag{40}$$

$$\mathscr{L}_{mom} = \lambda_{L1}||m - \hat{m}|| + \lambda_{gIoU}\mathscr{L}_{gIoU}(m, \hat{m}), \tag{41}$$

$$\mathscr{L}_{hd} = \lambda_{marg}\mathscr{L}_{marg} + \lambda_{cont}\text{RAC}(X_r^{\text{pos}}, X_r^{\text{neg}}) \tag{42}$$

$$\mathscr{L}_{marg} = \max(0, \Delta + S(x^{\text{low}}) - S(x^{\text{high}})), \tag{43}$$

其中 y 和 $\hat{y}$ 是前景或背景的眞值及其對應的預測，m 和 $\hat{m}$ 是眞值矩及其對應的預測，IoU 損失 $\mathscr{L}_{gIoU}(\cdot,\cdot)$ 來自之前的研究 (Rezatofighi et al. 2019),Δ 是邊距，$S(\cdot)$ 是顯著性分數估計器，x^{high} 和 x^{low} 分別是來自兩對高秩和低秩剪輯的視頻標記，$\text{CE}(\cdot,\cdot)$ 是交叉熵損失，$\text{RAC}(\cdot,\cdot)$ 是排名感知對比損失 (Hoffmann et al. 2022)，X_r^{pos} 和 X_r^{neg} 分別是比迭代索引排序高和低的樣本集合。

4 實驗

4.1 數據集

我們在四個流行的視頻時刻檢索和精彩片段檢測公共數據集上評估了我們的模型：QVHighlights (Lei, Berg, and Bansal 2021) (Gao et al. 2017) (Regneri et al. 2013) 和 TVSum (Song et al. 2015)。

由於數據集限制了提交的次數，在與模型的比較中，我們進行了多次實驗，並給出了最有可能獲得最佳結果的實驗結果。

因爲 QVHighlight (Lei, Berg, and Bansal 2021) 是目前唯一一個同時支持視頻時刻檢索和精彩片段檢測的數據集。因此，在消融實驗中，我們在 QVHighlight 數據集驗證集上進行了所有的實驗。我們每個實驗進行了五次，分別使用 1、23、456、7890 和 1,2345[1]作爲隨機種子，並給出了所有實驗結果的平均值和方差。

4.2 指標

我們採用了與之前研究相同的評估指標。具體來說，我們計算了 IoU 閾值 $\theta_{IoU} = 0.5$ 和0.7 的 Recall@1、$\theta_{IoU} = 0.5$ 和0.7 的 mAP（平均精度，mean average precision）以及一系列閾值 [0.5 : 0.05 : 0.95] 的 mAP，用於 QVHighlights 上的視頻時刻檢索。精彩片段檢測採用 mAP 和 HIT@1，其中正樣本定義爲顯著性得分爲 v。在 Charades-STA 和 TACoS 數據集上，我們使用 Recall@1 和 $\theta_{IoU} = \{0.3, 0.5, 0.7\}$ 和 mIoU 來衡量視頻時刻檢索性能。對於 TVSum，分別採用 mAP 和 Top-5 mAP。

4.3 實驗設置

在所有實驗中，我們採用 CLIP (Radford et al. 2021) 和 Slowfast (Feichtenhofer et al. 2019) 作爲提取器來提取視頻特徵，採用 CLIP (Radford et al. 2021) 提取文本特徵。在部分實驗中，採用 PANN (Kong et al. 2020) 提取音頻特徵。

默認情況下，我們的模型使用 AdamW 優化器 (Loshchilov 2017) 訓練了 200 個 epoch，學習率爲 1e-4，批次大小爲 32，隱藏維度爲 256，隊列長度爲 6,5536，動量（momentum）係數爲 0.995，提取係數（distillation coefficient）爲 0.4，卷積塊（Convolutional Blocks）層數爲 5，參考點（reference points）數爲 10，循環解碼器循環數爲 3。在 Charade-STA 數據集中，我們使用 4 層卷積塊。在 TACoS 數據集中，我們使用蒸餾係數爲 0.7。在 Charade-STA 數據集中，我們使用提取係數爲 0.3 訓練了 100 個 epoch。在 TVSum 數據集中，我們嘗試了每種方法並更改了每個超參數以獲得更好的結果。由於此數據集的特性，每次實驗中的超參數都不同。因此，我們沒有記錄它們。

4.4 與其他模型的比較

圖 6 直觀地展示了 LD-DETR 與其他模型的比較。表 1 報告了 LD-DETR 在 QVHighlight 數據集上聯合視頻時刻檢索和精彩片段檢測任務中的表現。表中給出的所有模

[1]爲方便閱讀，我們在本論文中使用「, 」作爲數字分位符，將數字以四位為一段進行分隔。

圖 6: 視頻時刻檢索和精彩片段檢測的可視化比較。

型均未使用預訓練。表中的模型根據使用的提取器分爲三類。我們的模型優於所有現有模型，甚至優於那些使用更多提取特徵的模型。表 2 報告了 LD-DETR 在 TACoS 數據集和 CharadesSTA 數據集上視頻時刻檢索的表現。得益於我們提出的方法，LD-DETR 模型在 QVHighlight、Charades-STA 和 TACoS 數據集上的表現優於最先進的模型。

表 3 報告了 LD-DETR 在 TV-Sum 數據集上進行高光檢測的表現。TV-Sum 數據集太小，每個類別只有 4 個訓練樣本和 1 個測試樣本。在訓練過程中，模型很快過擬合。即便如此，LD-DETR 模型在 TV-Sum 數據集上仍然取得了不錯的效果。

4.5 消融實驗

提取對齊的消融實驗 表 4 展示了提取對齊在多個模型上的表現。它展示了提取對齊作爲一種即插即用的方法，可以提高多個模型的性能。通過提取對齊方法，將映射到潛在空間的視頻和文本特徵對齊到同一個語義空間，並在對齊過程中考慮到不同訓練樣本中重疊的語義信息。

圖 7 和表 5 顯示，提取對齊方法能夠使模型取得較好的效果，同時不會佔用過多的 GPU 內存。可以看出，在使用更大的批次大小時，隨著批次大小的增大，模型的效果會

模型	視頻時刻檢索					精彩片段檢測	
	R1		mAP			>=Very Good	
	@0.5	@0.7	@0.5	@0.75	Avg.	mAP	HIT@1
Slowfast + CLIP (5.5 GB)							
BeautyThumb (Song et al. 2016)	-	-	-	-	-	14.36	22.88
DVSE (Liu et al. 2015)	-	-	-	-	-	18.75	21.79
MCN (Anne Hendricks et al. 2017)	11.41	2.72	24.94	8.22	10.67	-	-
CAL (Escorcia et al. 2019a)	25.49	11.54	23.40	7.65	9.89	-	-
XML (Lei et al. 2020)	41.83	30.35	44.63	31.73	32.14	34.49	55.25
XML$^+$ (Lei, Berg, and Bansal 2021)	46.69	33.46	47.89	34.67	34.90	35.38	55.06
Moment-DETR (Lei, Berg, and Bansal 2021)	52.89	33.02	54.82	29.40	30.73	35.69	55.60
Localizer (Yu et al. 2024)	54.50	36.50	-	-	32.30	-	-
UniVTG (Lin et al. 2023)	58.86	40.86	57.60	35.59	35.47	38.20	60.96
MomentDiff (Li et al. 2024)	57.42	39.66	54.02	35.73	35.95	-	-
VMRNet (繆翌, 張衛鋒, and 徐領 2024)	59.94	42.84	55.56	37.75	36.87	-	-
LLaViLo (Ma et al. 2023)	59.23	41.42	9.72	-	36.94	-	-
MH-DETR (Xu et al. 2024)	60.05	42.48	60.75	38.13	38.38	38.22	60.51
QD-DETR (Moon et al. 2023b)	62.40	44.98	62.52	39.88	39.86	38.94	62.40
CDIM (黎金宇 2024)	60.51	45.53	61.36	41.05	39.94	37.69	60.05
BM-DETR (Jung et al. 2023)	60.12	43.05	63.08	40.18	40.08	-	-
MESM (Liu et al. 2024c)	62.78	45.20	62.64	41.45	40.68	-	-
EaTR (Jang et al. 2023)	61.36	45.79	61.86	41.91	41.74	37.15	58.65
LMR (Liu et al. 2024a)	64.40	47.21	64.65	43.16	42.56	-	-
TR-DETR (Sun et al. 2024)	64.66	48.96	63.98	43.73	42.62	39.91	63.42
CG-DETR (Moon et al. 2023a)	65.43	48.38	64.51	42.77	42.86	40.33	**66.21**
CDNet (Ma et al. 2024)	67.74	49.55	63.82	42.30	42.76	39.84	66.52
UVCOM (Xiao et al. 2024)	63.55	47.47	63.37	42.67	43.18	39.74	64.20
SFABD (Huang et al. 2024)	-	-	62.38	44.39	43.79	-	-
LLMEPET (Jiang et al. 2024)	66.73	49.94	65.76	43.91	44.05	40.33	65.69
UniVTG$^+$ (Chen et al. 2024)	66.65	**52.19**	64.37	46.68	45.18	40.18	64.77
BAM-DETR (Lee and Byun 2023)	62.71	48.64	64.57	46.33	45.36	-	-
TaskWeave (Yang et al. 2024)	64.26	50.06	65.39	46.47	45.38	39.28	63.68
LD-DETR (我們的模型)	**66.80**	51.04	**67.61**	**46.99**	**46.41**	**40.51**	65.11
Slowfast + CLIP + PANN (11.7 GB)							
UMT (Liu et al. 2022b)	56.23	41.18	53.38	37.01	36.12	38.18	59.99
VCSJT (Zhou et al. 2024)	59.14	42.02	55.76	37.79	36.37	38.59	62.45
MIM (Li et al. 2023)	59.99	41.50	55.85	36.84	36.45	38.96	62.39
LSJT (Wang et al. 2024)	60.51	41.50	56.33	36.70	36.66	39.13	61.22
MomentDiff (Li et al. 2024)	58.21	41.48	54.57	37.21	36.84	-	-
QD-DETR (Moon et al. 2023b)	63.06	45.10	63.04	40.10	40.19	39.04	62.87
UVCOM (Xiao et al. 2024)	63.81	48.70	64.47	44.01	43.27	39.79	64.79
LD-DETR (我們的模型)	**65.76**	**50.71**	**66.06**	**46.62**	**45.85**	**41.00**	**67.06**
CLIP$^+$ (233.7 GB)							
R^2-Tuning (Liu et al. 2024b)	68.03	49.35	69.04	47.56	46.17	40.75	64.20

表 1: QVHighlights 測試集上的聯合視頻時刻檢索和高光檢測結果。該表按使用的提取器對模型進行分類，並標明提取器和提取的特徵大小。每列中測試集上每類特徵中的最佳結果以**粗體**突出顯示，第二佳結果以下劃線 突出顯示。灰色顯示的模型僅報告其在論文中的驗證集上的結果。

越來越好。但是，批次大小越大，模型佔用的 GPU 內存也會越大，並且隨著批次大小的增大，模型收斂的難度也會越大。具體來說，在批次大小 = 1,024 時，我們進行了 5 次實驗，其中只有一次模型成功收斂。而使用提取對齊時，隨著隊列長度的增加，模型的效果不斷變好，而佔用的 GPU 內存卻小了很多。

模型	TACoS				Charades-STA			
	R1@0.3	R1@0.5	R1@0.7	mIoU	R1@0.3	R1@0.5	R1@0.7	mIoU
CTRL (Gao et al. 2017)	18.32	13.30	-	-	-	23.63	8.89	-
ABLR (Yuan, Mei, and Zhu 2019)	19.50	9.40	-	13.40	-	-	-	-
SM-RL (Wang, Huang, and Wang 2019)	-	20.25	15.95	-	-	24.36	11.17	-
TGN (Chen et al. 2018)	21.77	18.90	-	-	-	-	-	-
SAP (Chen and Jiang 2019)	-	18.24	-	-	-	27.42	13.36	-
MIM† (Li et al. 2023)	-	-	-	-	-	43.92	25.89	-
MAN (Zhang et al. 2019)	-	-	-	-	-	46.53	22.72	-
FMAN (蔣尋 et al. 2023)	-	-	-	-	-	51.40	25.05	38.23
LSJT† (Wang et al. 2024)	-	-	-	-	-	44.62	25.13	-
DRN (Zeng et al. 2020)	-	23.17	-	-	-	45.40	26.40	-
UMT† (Liu et al. 2022b)	-	-	-	-	-	48.31	29.25	-
VCSJT† (Zhou et al. 2024)	-	-	-	-	-	51.21	30.22	-
SFABD (Huang et al. 2024)	-	-	-	-	-	-	30.51	-
BPNet (Xiao et al. 2021)	25.96	20.96	14.08	19.53	65.48	50.75	31.64	46.34
M-DETR (Lei, Berg, and Bansal 2021)	-	-	-	-	-	53.63	31.37	-
SCDM (Yuan et al. 2019b)	26.11	21.17	-	-	-	54.44	31.37	-
DCL (Nan et al. 2021)	38.84	29.07	19.05	28.26	67.63	50.24	32.88	48.02
HUAL (Ji et al. 2023)	-	-	-	-	70.40	52.69	28.90	48.11
VSLNet (Zhang et al. 2020a)	29.61	24.27	20.03	24.11	70.46	54.19	35.22	50.02
2D-TAN (Zhang et al. 2020b)	37.29	25.32	-	-	-	39.70	23.31	-
MMN (Wang et al. 2022b)	39.24	26.17	-	-	-	47.31	27.28	-
CrossGraphAlign (陳卓 et al. 2020)	39.80	26.40	-	-	-	-	-	-
CBLN (Liu et al. 2021)	38.98	27.65	-	-	-	61.13	38.22	-
CPNet (Li, Guo, and Wang 2021)	42.61	28.29	-	28.69	-	60.27	38.74	-
FVMR (Gao et al. 2021b)	41.48	29.12	-	-	-	59.46	35.48	-
SimVTP (Ma et al. 2022)	43.10	30.30	-	-	-	44.70	26.30	-
RaNet (Gao et al. 2021a)	43.34	33.54	-	-	-	60.40	39.65	-
MomentDiff (Li et al. 2024)	44.78	33.68	-	-	-	55.57	32.42	-
LLaViLo (Ma et al. 2023)	-	-	-	-	-	55.72	33.43	-
TaskWeave (Yang et al. 2024)	-	-	-	-	-	56.51	33.66	-
QD-DETR† (Moon et al. 2023b)	-	-	-	-	-	55.51	34.17	-
LMR (Liu et al. 2024a)	-	-	-	-	-	55.91	35.19	-
VLG-Net (Soldan et al. 2021)	45.46	34.19	-	-	-	-	-	-
GVL (Wang et al. 2023)	45.92	34.57	-	32.48	-	-	-	-
TR-DETR (Sun et al. 2024)	-	-	-	-	-	57.61	33.52	-
UniVTG (Lin et al. 2023)	51.44	34.97	17.35	33.60	70.81	58.01	35.65	50.10
UniVTG$^+$ (Chen et al. 2024)	-	-	-	-	68.06	57.18	36.05	-
BM-DETR (Jung et al. 2023)	50.31	35.42	-	-	-	59.48	38.33	-
SFEN (楊金福 et al. 2022)	47.30	36.10	-	-	-	-	-	-
MATN (Zhang et al. 2021)	48.79	37.57	-	-	-	-	-	-
這個模型沒有名字 (Panta et al. 2024)	49.77	37.99	-	-	-	-	-	-
CDNet (Ma et al. 2024)	54.11	35.35	20.34	33.76	71.25	58.09	36.53	-
MS-DETR (Jing et al. 2023)	47.66	37.36	25.81	35.09	68.68	57.72	37.40	50.12
CG-DETR (Moon et al. 2023a)	52.23	39.61	22.23	36.48	70.43	58.44	36.34	50.13
UVCOM (Xiao et al. 2024)	-	36.39	23.32	-	-	59.25	36.64	-
R^2-Tuning (Liu et al. 2024b)	49.71	38.72	25.12	35.92	70.91	59.78	37.02	50.86
LLMEPET (Jiang et al. 2024)	52.73	-	22.78	36.55	70.91	-	36.49	50.25
UnLoc (Yan et al. 2023)	-	-	-	-	-	60.80	38.40	-
MESM (Liu et al. 2024c)	52.69	39.52	-	36.94	-	61.24	38.04	-
MCMN (Han et al. 2023)	50.24	36.78	-	-	-	**62.69**	41.38	-
BAM-DETR (Lee and Byun 2023)	56.69	41.54	**26.77**	39.31	72.93	59.95	39.38	52.33
LD-DETR (我們的模型)	**57.61**	**44.31**	26.24	**40.30**	**73.06**	62.28	**42.23**	**53.14**

表 2: TACoS 和 Charades-STA 上的視頻時刻檢索結果。每列中的最佳結果以**粗體**突出顯示，第二佳結果以下劃線 突出顯示。† 表示使用音頻模態進行訓練。

模型	VT	VU	GA	MS	PK	PR	FM	BK	BT	DS	**Avg.**
sLSTM (Zhang et al. 2016)	41.1	46.2	46.3	47.7	44.8	46.1	45.2	40.6	47.1	45.5	45.1
SG (Yuan et al. 2019a)	42.3	47.2	47.5	48.9	45.6	47.3	46.4	41.7	48.3	46.6	46.2
VESD (Cai et al. 2018)	44.7	49.3	49.6	50.3	47.8	48.5	48.7	44.1	49.2	48.8	48.1
LIM-S (Xiong et al. 2019)	55.9	42.9	61.2	54.0	60.3	47.5	43.2	66.3	69.1	62.6	56.3
Trailer (Wang et al. 2020)	61.3	54.6	65.7	60.8	59.1	70.1	58.2	64.7	65.6	68.1	62.8
MINI-Net† (Hong et al. 2020)	80.6	68.3	78.2	81.8	78.1	65.8	75.8	75.0	80.2	65.5	73.2
SL-Module (Xu et al. 2021)	86.5	68.7	74.9	86.2	79.0	63.2	58.9	72.6	78.9	64.0	73.3
SA (Badamdorj et al. 2021)	83.4	64.7	84.4	86.5	70.3	67.5	66.9	68.1	95.0	60.8	74.8
SA$^+$† (Badamdorj et al. 2021)	83.7	57.3	78.5	86.1	80.1	69.2	70.0	73.0	97.4	67.5	76.3
Joint-VA† (Badamdorj et al. 2021)	83.7	57.3	78.5	86.1	80.1	69.2	70.0	73.0	97.4	67.5	76.3
TCG† (Ye et al. 2021)	85.0	71.4	81.9	78.6	80.2	75.5	71.6	77.3	78.6	68.1	76.8
PLD-VHD (Wei et al. 2022)	84.5	80.9	70.3	72.5	76.4	87.2	71.9	74.0	74.4	79.1	77.1
UniVTG (Lin et al. 2023)	83.9	85.1	89.0	80.1	84.6	87.0	70.9	91.7	73.5	69.3	81.0
MH-DETR (Xu et al. 2024)	86.1	79.4	84.3	85.8	81.2	83.9	74.3	82.7	86.5	71.6	81.6
MIM† (Li et al. 2023)	84.4	85.8	91.3	73.9	83.1	87.1	80.1	78.2	80.3	79.6	82.4
UMT† (Liu et al. 2022b)	87.5	81.5	88.2	78.8	81.4	87.0	76.0	86.9	84.4	79.6	83.1
VCSJT† (Zhou et al. 2024)	87.5	80.7	88.6	76.6	83.6	91.0	77.6	93.3	88.9	80.0	84.8
QD-DETR (Moon et al. 2023b)	88.2	87.4	85.6	85.0	85.8	86.9	76.4	91.3	89.2	73.7	85.0
LD-DETR (我們的模型)	83.1	90.3	91.5	82.5	87.7	83.5	79.0	90.4	86.1	77.3	85.1
UVCOM (Xiao et al. 2024)	87.6	91.6	91.4	86.7	86.9	86.9	76.9	92.3	87.4	75.6	86.3
QD-DETR† (Moon et al. 2023b)	87.6	91.7	90.2	88.3	84.1	88.3	78.7	91.2	87.8	77.7	86.6
CG-DETR (Moon et al. 2023a)	86.9	88.8	94.8	87.7	86.7	89.6	74.8	93.3	89.2	75.9	86.8
TR-DETR† (Sun et al. 2024)	89.3	93.0	94.3	85.1	88.0	88.6	80.4	91.3	89.5	81.6	88.1
TaskWeave (Yang et al. 2024)	88.2	90.8	93.3	87.5	87.0	82.0	80.9	92.9	89.5	81.2	87.3
LLMEPET (Jiang et al. 2024)	90.8	92.0	93.8	81.5	87.5	86.0	79.6	96.2	88.0	79.0	87.4
TR-DETR (Sun et al. 2024)	89.3	93.0	94.3	85.1	88.0	88.6	80.4	91.3	89.5	81.6	88.1
CDIM (黎金宇 2024)	85.5	95.8	90.3	90.0	88.4	88.1	79.2	97.1	88.0	80.5	88.3

表 3: TV-Sum 上的精彩片段檢測結果。† 表示使用音頻模態進行訓練。

設置	視頻時刻檢索					精彩片段檢測	
	R1		mAP			>=Very Good	
	@0.5	@0.7	@0.5	@0.75	Avg.	mAP	HIT@1
M-DETR (Lei, Berg, and Bansal 2021)	$53.63_{\pm 1.59}$	$35.78_{\pm 1.25}$	$54.99_{\pm 1.33}$	$31.28_{\pm 1.07}$	$32.09_{\pm 1.01}$	$36.41_{\pm 0.40}$	$56.92_{\pm 0.98}$
M-DETR + **提取對齊**	$\mathbf{55.56}_{\pm \mathbf{1.01}}$	$\mathbf{36.72}_{\pm \mathbf{1.43}}$	$\mathbf{56.42}_{\pm \mathbf{0.76}}$	$\mathbf{31.43}_{\pm \mathbf{0.99}}$	$\mathbf{32.66}_{\pm \mathbf{0.73}}$	$\mathbf{37.16}_{\pm \mathbf{0.36}}$	$\mathbf{58.62}_{\pm \mathbf{1.19}}$
BM-DETR (Jung et al. 2023)	$60.90_{\pm 0.79}$	$44.10_{\pm 1.43}$	$60.91_{\pm 0.96}$	$39.03_{\pm 1.46}$	$\mathbf{38.93}_{\pm \mathbf{0.98}}$	-	-
BM-DETR + **提取對齊**	$\mathbf{61.04}_{\pm \mathbf{1.02}}$	$\mathbf{44.23}_{\pm \mathbf{1.04}}$	$\mathbf{61.53}_{\pm \mathbf{0.98}}$	$\mathbf{39.12}_{\pm \mathbf{1.45}}$	$38.68_{\pm 1.09}$	-	-
QD-DETR (Moon et al. 2023b)	$61.98_{\pm 0.55}$	$47.30_{\pm 0.69}$	$62.03_{\pm 0.43}$	$41.96_{\pm 0.66}$	$41.42_{\pm 0.28}$	$38.92_{\pm 0.30}$	$62.05_{\pm 0.86}$
QD-DETR + **提取對齊**	$\mathbf{64.10}_{\pm \mathbf{0.42}}$	$\mathbf{48.53}_{\pm \mathbf{0.70}}$	$\mathbf{63.71}_{\pm \mathbf{0.37}}$	$\mathbf{43.55}_{\pm \mathbf{0.59}}$	$\mathbf{42.80}_{\pm \mathbf{0.46}}$	$\mathbf{39.58}_{\pm \mathbf{0.40}}$	$\mathbf{63.21}_{\pm \mathbf{1.05}}$
CG-DETR (Moon et al. 2023a)	$65.92_{\pm 0.22}$	$50.44_{\pm 0.54}$	$65.50_{\pm 0.22}$	$45.62_{\pm 0.68}$	$44.76_{\pm 0.26}$	$40.34_{\pm 0.20}$	$65.12_{\pm 0.64}$
CG-DETR + **提取對齊**	$\mathbf{66.11}_{\pm \mathbf{0.79}}$	$\mathbf{51.18}_{\pm \mathbf{0.89}}$	$\mathbf{65.65}_{\pm \mathbf{0.61}}$	$\mathbf{46.12}_{\pm \mathbf{0.74}}$	$\mathbf{45.23}_{\pm \mathbf{0.61}}$	$\mathbf{40.50}_{\pm \mathbf{0.23}}$	$\mathbf{65.92}_{\pm \mathbf{1.04}}$
TR-DETR (Sun et al. 2024)	$66.56_{\pm 1.06}$	$50.13_{\pm 0.89}$	$65.70_{\pm 0.79}$	$45.10_{\pm 0.78}$	$44.33_{\pm 0.51}$	$40.88_{\pm 0.19}$	$65.54_{\pm 0.45}$
TR-DETR − LGMA	$62.28_{\pm 1.08}$	$46.99_{\pm 1.21}$	$62.38_{\pm 0.72}$	$62.16_{\pm 1.41}$	$41.56_{\pm 1.12}$	$39.16_{\pm 0.32}$	$62.13_{\pm 0.46}$
TR-DETR + **提取對齊**	$66.31_{\pm 0.60}$	$49.92_{\pm 0.73}$	$65.33_{\pm 0.74}$	$44.14_{\pm 0.81}$	$43.60_{\pm 0.94}$	$\mathbf{40.89}_{\pm \mathbf{0.26}}$	$\mathbf{65.79}_{\pm \mathbf{1.02}}$
TR-DETR − LGMA + **提取對齊**	$\mathbf{67.14}_{\pm \mathbf{0.43}}$	$\mathbf{51.17}_{\pm \mathbf{0.37}}$	$\mathbf{66.21}_{\pm \mathbf{0.24}}$	$\mathbf{45.57}_{\pm \mathbf{0.42}}$	$\mathbf{44.89}_{\pm \mathbf{0.32}}$	$40.77_{\pm 0.31}$	$65.33_{\pm 0.71}$
UVCOM (Xiao et al. 2024)	$64.36_{\pm 0.44}$	$50.21_{\pm 1.06}$	$63.99_{\pm 0.27}$	$45.52_{\pm 0.68}$	$44.77_{\pm 0.53}$	$39.85_{\pm 0.21}$	$63.82_{\pm 1.05}$
UVCOM + **提取對齊**	$\mathbf{66.39}_{\pm \mathbf{0.37}}$	$\mathbf{51.54}_{\pm \mathbf{0.22}}$	$\mathbf{65.24}_{\pm \mathbf{0.49}}$	$\mathbf{46.18}_{\pm \mathbf{0.38}}$	$\mathbf{45.39}_{\pm \mathbf{0.29}}$	$\mathbf{40.72}_{\pm \mathbf{0.14}}$	$\mathbf{65.20}_{\pm \mathbf{0.73}}$

表 4: 提取對齊作爲一種即插即用的方法，可以幫助多個模型獲得更好的結果。LGMA 是局部-全局多模態比對 (Local-Global Multi-Modal Alignment) 的縮寫，是 TR-DETR (Sun et al. 2024) 模型中使用的一種方法。每列中每個基準模型上的最佳結果以**粗體**突出顯示。

批次大小/隊列長度	GPU 內存	視頻時刻檢索					精彩片段檢測	
		R1		mAP			>=Very Good	
		@0.5	@0.7	@0.5	@0.75	Avg.	mAP	HIT@1
更大的批次大小								
bs = 32, ql = 0	1,350	$66.19_{\pm 1.41}$	$49.74_{\pm 1.70}$	$65.27_{\pm 1.11}$	$44.12_{\pm 1.41}$	$43.81_{\pm 1.19}$	$40.72_{\pm 0.17}$	$65.19_{\pm 1.23}$
bs = 64, ql = 0	2,140	$66.89_{\pm 0.75}$	$50.07_{\pm 0.82}$	$65.73_{\pm 0.39}$	$44.53_{\pm 0.54}$	$43.98_{\pm 0.50}$	$40.99_{\pm 0.29}$	$65.86_{\pm 0.84}$
bs = 128, ql = 0	3,328	$66.81_{\pm 0.70}$	$50.93_{\pm 0.46}$	$65.93_{\pm 0.46}$	$45.00_{\pm 0.43}$	$44.53_{\pm 0.33}$	$40.92_{\pm 0.20}$	$65.77_{\pm 1.26}$
bs = 256, ql = 0	5,772	$66.62_{\pm 0.59}$	$49.56_{\pm 0.82}$	$65.27_{\pm 0.51}$	$43.46_{\pm 0.66}$	$43.22_{\pm 0.83}$	$40.76_{\pm 0.16}$	$65.38_{\pm 0.52}$
bs = 512, ql = 0	10,530	$65.20_{\pm 1.92}$	$47.38_{\pm 2.31}$	$63.49_{\pm 1.85}$	$41.35_{\pm 2.33}$	$40.90_{\pm 2.00}$	$40.58_{\pm 0.17}$	$64.94_{\pm 1.16}$
bs = 1024, ql = 0	20,956	$19.15_{\pm 13.66}$	$8.70_{\pm 8.05}$	$26.29_{\pm 10.96}$	$9.45_{\pm 6.48}$	$11.73_{\pm 6.34}$	$27.13_{\pm 5.20}$	$37.18_{\pm 10.86}$
bs = 2048, ql = 0	內存不足	-	-	-	-	-	-	-
提取對齊 (我們的方法)								
bs = 32, ql = 0	1,312	$66.53_{\pm 0.96}$	$49.90_{\pm 0.31}$	$65.70_{\pm 0.89}$	$44.57_{\pm 0.48}$	$44.07_{\pm 0.13}$	$40.52_{\pm 0.24}$	$64.73_{\pm 0.95}$
bs = 32, ql = 96	1,318	$66.53_{\pm 0.32}$	$49.34_{\pm 0.33}$	$65.61_{\pm 0.85}$	$44.13_{\pm 0.35}$	$43.77_{\pm 0.49}$	$40.50_{\pm 0.16}$	$64.66_{\pm 0.24}$
bs = 32, ql = 480	1,236	$67.11_{\pm 1.49}$	$50.48_{\pm 1.03}$	**$66.28_{\pm 1.03}$**	$45.10_{\pm 0.51}$	$44.18_{\pm 0.41}$	$40.92_{\pm 0.23}$	$65.91_{\pm 0.97}$
bs = 32, ql = 2016	1,458	$66.76_{\pm 0.57}$	$50.48_{\pm 0.40}$	$65.73_{\pm 0.47}$	$44.83_{\pm 0.68}$	$44.22_{\pm 0.57}$	$40.81_{\pm 0.39}$	$65.09_{\pm 0.70}$
bs = 32, ql = 8160	1,364	$66.74_{\pm 0.84}$	**$50.81_{\pm 0.50}$**	$66.07_{\pm 0.42}$	**$45.30_{\pm 0.52}$**	$44.53_{\pm 0.34}$	$40.76_{\pm 0.19}$	**$65.95_{\pm 0.42}$**
bs = 32, ql = 3,2736	1,620	$67.00_{\pm 0.75}$	$50.71_{\pm 1.10}$	$66.10_{\pm 0.67}$	$45.06_{\pm 0.99}$	$44.51_{\pm 0.80}$	**$41.00_{\pm 0.21}$**	$65.64_{\pm 0.53}$
bs = 32, ql = 13,1040	2,440	**$67.14_{\pm 0.59}$**	$50.63_{\pm 0.48}$	$65.91_{\pm 0.32}$	$45.21_{\pm 0.60}$	**$44.67_{\pm 0.30}$**	**$41.00_{\pm 0.23}$**	$65.86_{\pm 0.82}$

表 5: 與使用更大的批次大小相比，使用提取對齊可以節省 GPU 內存並獲得更好的結果。下表顯示了使用不同批次大小或隊列長度、使用更大的批次大小和使用提煉對齊時的 GPU 內存使用情況和訓練結果。「 bs 」表示批次大小，「 ql 」表示隊列長度。實驗在具有 2,4564MiB GPU 內存的 NVIDIA GeForce RTX 4090 上使用 TR-DETR (Sun et al. 2024) 作爲基準模型進行。每列中的最佳結果以**粗體**突出顯示。此表顯示了與圖 7 相同的一組實驗。

提取係數	視頻時刻檢索					精彩片段檢測	
	R1		mAP			>=Very Good	
	@0.5	@0.7	@0.5	@0.75	Avg.	mAP	HIT@1
0.0	$66.28_{\pm 1.38}$	$49.91_{\pm 0.44}$	$65.64_{\pm 0.74}$	$44.60_{\pm 0.68}$	$44.23_{\pm 0.27}$	$40.81_{\pm 0.21}$	$65.03_{\pm 0.78}$
0.1	$67.20_{\pm 0.64}$	$50.50_{\pm 0.78}$	$66.20_{\pm 0.40}$	$44.83_{\pm 0.64}$	$44.50_{\pm 0.60}$	$40.85_{\pm 0.20}$	$65.58_{\pm 0.89}$
0.2	$67.15_{\pm 0.99}$	$50.50_{\pm 0.26}$	$66.28_{\pm 0.61}$	$45.24_{\pm 0.39}$	$44.63_{\pm 0.40}$	$40.98_{\pm 0.18}$	$66.00_{\pm 0.65}$
0.3	$67.07_{\pm 0.59}$	$50.62_{\pm 0.42}$	$66.19_{\pm 0.54}$	$45.29_{\pm 0.52}$	$44.61_{\pm 0.49}$	**$41.14_{\pm 0.16}$**	$65.78_{\pm 1.26}$
0.4	**$67.59_{\pm 0.71}$**	**$50.89_{\pm 0.82}$**	**$66.48_{\pm 0.37}$**	**$45.50_{\pm 0.58}$**	**$44.86_{\pm 0.50}$**	$41.09_{\pm 0.22}$	**$66.32_{\pm 0.47}$**
0.5	$66.76_{\pm 0.89}$	$50.46_{\pm 0.62}$	$66.16_{\pm 0.54}$	$45.06_{\pm 0.50}$	$44.66_{\pm 0.28}$	$40.83_{\pm 0.37}$	$65.56_{\pm 0.62}$
0.6	$66.61_{\pm 0.43}$	$50.55_{\pm 0.59}$	$65.89_{\pm 0.54}$	$45.32_{\pm 0.56}$	$44.68_{\pm 0.49}$	$40.65_{\pm 0.20}$	$64.55_{\pm 0.67}$
0.7	$66.59_{\pm 0.28}$	$50.52_{\pm 0.81}$	$65.71_{\pm 0.22}$	$45.14_{\pm 0.49}$	$44.63_{\pm 0.32}$	$41.05_{\pm 0.19}$	$65.43_{\pm 0.42}$
0.8	$65.72_{\pm 0.94}$	$49.55_{\pm 1.09}$	$64.87_{\pm 0.57}$	$44.05_{\pm 1.51}$	$43.62_{\pm 1.09}$	$40.60_{\pm 0.13}$	$64.80_{\pm 0.50}$
0.9	$65.25_{\pm 0.70}$	$49.47_{\pm 0.33}$	$64.75_{\pm 0.22}$	$44.17_{\pm 0.36}$	$43.69_{\pm 0.33}$	$40.23_{\pm 0.34}$	$63.88_{\pm 1.25}$
1.0	$62.83_{\pm 0.82}$	$47.26_{\pm 0.85}$	$63.08_{\pm 0.49}$	$42.54_{\pm 0.82}$	$42.09_{\pm 0.58}$	$39.37_{\pm 0.21}$	$61.87_{\pm 0.39}$

表 6: 該表顯示了提取對齊中提取係數對多模態比齊的影響。每列中每個基線上的最佳結果以**粗體**突出顯示。此表顯示了與圖 8 相同的一組實驗。

圖 8 和表 6 展示了提取對齊中提取係數對多模態對齊的影響。其中，當提取係數 = 0.0 時，相當於不啓用提取。可以看出，當使用一定程度的提取時，不同訓練樣本之間的重疊語義信息被多模態對齊考慮在內，模型的效果變得更好。

卷積融合器的消融實驗 表 7 展示了卷積融合器中卷積層數量的影響。視頻中的運動信息往往蘊含在若干個局部片段中，卷積層可以有效的捕捉視頻中的局部信息。與 UVCOM (Xiao et al. 2024) (a-f) 相比，我們的方法 (g-m) 可以更好地利用卷積層提取局部特徵的能力。在 UVCOM 中，隨著卷積層數量的增加，模型的性能並沒有變好。但在我們的方法中，隨著卷積層數量的增加，模型的性能變得更好。與不使用卷積層 (g) 相比，僅添加一個卷積層 (h) 可以大大提升模型效果。隨著加入的卷積層數量

設置	# Conv	視頻時刻檢索					精彩片段檢測	
		R1		mAP			>=Very Good	
		@0.5	@0.7	@0.5	@0.75	Avg.	mAP	HIT@1
UVCOM (Xiao et al. 2024)								
(a)	0	$63.68_{\pm 0.43}$	$49.29_{\pm 0.86}$	$63.50_{\pm 0.46}$	$43.81_{\pm 0.44}$	$43.40_{\pm 0.41}$	$39.40_{\pm 0.13}$	$63.22_{\pm 0.46}$
(b)	1	$64.36_{\pm 0.44}$	$50.21_{\pm 1.06}$	$63.99_{\pm 0.27}$	$45.52_{\pm 0.68}$	$44.72_{\pm 0.53}$	$39.85_{\pm 0.21}$	$63.82_{\pm 1.05}$
(c)	2	$61.68_{\pm 0.59}$	$47.51_{\pm 0.54}$	$61.72_{\pm 0.74}$	$42.66_{\pm 0.53}$	$42.40_{\pm 0.47}$	$38.64_{\pm 0.43}$	$60.90_{\pm 0.81}$
(d)	4	$61.89_{\pm 0.63}$	$47.54_{\pm 0.85}$	$61.63_{\pm 0.82}$	$42.46_{\pm 0.78}$	$42.46_{\pm 0.81}$	$38.52_{\pm 0.13}$	$60.54_{\pm 0.75}$
(e)	8	$60.87_{\pm 0.23}$	$46.95_{\pm 0.63}$	$61.41_{\pm 0.32}$	$42.33_{\pm 0.50}$	$42.28_{\pm 0.23}$	$38.39_{\pm 0.23}$	$60.22_{\pm 0.92}$
(f)	16	$61.46_{\pm 0.60}$	$47.13_{\pm 0.42}$	$61.76_{\pm 0.25}$	$43.07_{\pm 0.62}$	$42.57_{\pm 0.33}$	$38.19_{\pm 0.23}$	$59.80_{\pm 0.77}$
LD-DETR (我們的模型)								
(g)	0	$67.07_{\pm 1.54}$	$50.96_{\pm 2.17}$	$65.79_{\pm 1.72}$	$45.17_{\pm 2.10}$	$44.18_{\pm 1.91}$	$41.42_{\pm 0.09}$	$66.30_{\pm 0.29}$
(h)	1	$68.14_{\pm 0.26}$	$52.01_{\pm 1.28}$	$66.70_{\pm 0.38}$	$46.24_{\pm 1.18}$	$45.71_{\pm 0.96}$	$41.70_{\pm 0.20}$	$\mathbf{67.41}_{\pm \mathbf{0.95}}$
(i)	2	$68.10_{\pm 0.63}$	$52.04_{\pm 0.45}$	$67.49_{\pm 0.78}$	$47.01_{\pm 0.25}$	$46.73_{\pm 0.32}$	$41.65_{\pm 0.24}$	$66.62_{\pm 0.62}$
(j)	4	$68.77_{\pm 0.40}$	$52.88_{\pm 0.66}$	$68.00_{\pm 0.38}$	$47.55_{\pm 0.66}$	$47.21_{\pm 0.46}$	$\mathbf{41.83}_{\pm \mathbf{0.15}}$	$67.30_{\pm 1.33}$
(k)	8	$68.52_{\pm 0.90}$	$52.22_{\pm 1.06}$	$67.70_{\pm 0.62}$	$67.14_{\pm 0.90}$	$47.28_{\pm 0.66}$	$41.55_{\pm 0.24}$	$66.41_{\pm 0.61}$
(l)	10	$\mathbf{69.01}_{\pm \mathbf{1.09}}$	$\mathbf{53.19}_{\pm \mathbf{0.38}}$	$\mathbf{68.43}_{\pm \mathbf{0.83}}$	$\mathbf{48.25}_{\pm \mathbf{0.59}}$	$\mathbf{47.93}_{\pm \mathbf{0.39}}$	$41.66_{\pm 0.15}$	$66.80_{\pm 0.96}$
(m)	16	$68.63_{\pm 0.97}$	$52.89_{\pm 0.62}$	$68.10_{\pm 0.64}$	$47.56_{\pm 0.36}$	$47.24_{\pm 0.46}$	$41.79_{\pm 0.15}$	$67.33_{\pm 0.42}$

表 7: 該表顯示了卷積層數量對模型性能的影響。「 # Conv 」表示卷積層的數量。每列中每個基線上的最佳結果以**粗體**突出顯示。可以注意到，在 LD-DETR 上，藉助卷積融合器，僅添加一個卷積層就可以顯著提高模型性能。隨著卷積層數量的增加，模型的性能會越來越好，直到達到極限。然而，在 UVCOM (Xiao et al. 2024) 上，添加卷積層並不能幫助模型表現得更好。

設置	位置	視頻時刻檢索					精彩片段檢測	
		R1		mAP			>=Very Good	
		@0.5	@0.7	@0.5	@0.75	Avg.	mAP	HIT@1
(a)	V2TEx 之前	$58.72_{\pm 2.48}$	$42.43_{\pm 3.08}$	$58.45_{\pm 2.87}$	$37.44_{\pm 2.88}$	$37.14_{\pm 2.84}$	$38.20_{\pm 0.62}$	$60.39_{\pm 1.60}$
(b)	V2TEx & T2VEn 之間	$66.64_{\pm 1.21}$	$49.61_{\pm 1.43}$	$65.84_{\pm 1.00}$	$43.99_{\pm 1.19}$	$43.51_{\pm 1.15}$	$41.20_{\pm 0.19}$	$66.36_{\pm 0.48}$
(c)	T2VEn & TrEn1 之間	$68.87_{\pm 0.68}$	$52.96_{\pm 0.32}$	$68.07_{\pm 0.56}$	$47.54_{\pm 0.41}$	$47.51_{\pm 0.39}$	$\mathbf{41.81}_{\pm \mathbf{0.13}}$	$66.79_{\pm 0.50}$
(d)	TrEn1 & TrEn2 之間	$\mathbf{69.01}_{\pm \mathbf{1.09}}$	$\mathbf{53.19}_{\pm \mathbf{0.38}}$	$\mathbf{68.43}_{\pm \mathbf{0.83}}$	$\mathbf{48.25}_{\pm \mathbf{0.59}}$	$\mathbf{47.93}_{\pm \mathbf{0.39}}$	$41.66_{\pm 0.15}$	$\mathbf{66.80}_{\pm \mathbf{0.96}}$
(e)	TrEn2 之後	$67.62_{\pm 0.96}$	$52.23_{\pm 0.53}$	$67.17_{\pm 0.17}$	$46.85_{\pm 0.23}$	$46.57_{\pm 0.19}$	$41.56_{\pm 0.28}$	$66.26_{\pm 0.52}$

表 8: 該表顯示了卷積塊在卷積融合器中放置在不同位置時模型的性能。可以注意到，當卷積塊放置在 Transformer 編碼器 1 和 Transformer 編碼器 2 之間時，模型性能最佳。在該表中，「 V2TEx 」表示 V2T 提取器，「 T2VEn 」表示 T2V 編碼器，「 TrEn1 」表示 Transformer 編碼器 1，「 TrEn2 」表示 Transformer 編碼器 2。每列中的最佳結果以**粗體**突出顯示。

的增加，模型能夠更有效地捕捉視頻中的局部信息，效果逐漸提升直至達到峰值 (l)。

表 8 和表 9 展示了卷積融合器上的消融實驗。無論改變方法的順序還是刪除其中任何一個，模型的性能都會下降。

循環解碼器的消融實驗 圖 1 展示了循環解碼器使視頻時刻檢索更加準確。它可視化了每次循環時循環解碼器輸出對應的視頻時刻檢索結果。隨著循環次數的增加，結果越來越接近眞實值。圖 9 、表 10 和表 11 展示了循環解碼器在多個模型上的表現。它展示了循環解碼器作爲一種即插即用的方法，可以提高多個模型的性能。通過循環解碼器方法，多模態信息被更充分地解碼。使用循環解碼器後，模型的性能得到了很大的提高。然而，當使用相同尺寸的更大的解碼器時，模型很快就會過擬合。我們注意到，循環解碼器在 UVCOM (Xiao et al. 2024) 上的效果不如在其他模型上明顯。我們推測這是因爲它的雙分支模態內聚合 (Dual Branches Intra-Modality Aggregation) 影響了我們方法的性能。當我們刪除片段文本對齊 (Clip-Text Alignment，CTA) 方法時，隨著循環次數的增加，性能會變得更好。但是當我們刪除插槽注意力 (Slot Attention，SA) 方法時，模型的性能會變得更好。

設置	V2TEx	T2VEn	TrEn1	ConBl	TrEn2	視頻時刻檢索					精彩片段檢測	
						R1		mAP			>=Very Good	
						@0.5	@0.7	@0.5	@0.75	Avg.	mAP	HIT@1
(a)		✓	✓	✓	✓	$63.82_{\pm 0.66}$	$49.24_{\pm 0.77}$	$63.94_{\pm 0.11}$	$44.65_{\pm 0.28}$	$44.28_{\pm 0.26}$	$39.95_{\pm 0.27}$	$62.92_{\pm 1.02}$
(b)	✓		✓	✓	✓	$68.17_{\pm 0.88}$	$52.41_{\pm 0.96}$	$67.11_{\pm 0.73}$	$46.74_{\pm 0.96}$	$46.53_{\pm 0.50}$	$41.50_{\pm 0.25}$	$66.27_{\pm 0.95}$
(c)	✓	✓		✓	✓	$68.09_{\pm 0.65}$	$52.61_{\pm 0.84}$	$67.53_{\pm 0.49}$	$47.37_{\pm 0.42}$	$47.03_{\pm 0.17}$	$41.43_{\pm 0.15}$	$66.62_{\pm 0.43}$
(d)	✓	✓	✓		✓	$67.07_{\pm 1.54}$	$50.96_{\pm 2.17}$	$65.79_{\pm 1.72}$	$45.17_{\pm 2.10}$	$44.18_{\pm 1.91}$	$41.42_{\pm 0.09}$	$66.30_{\pm 0.29}$
(e)	✓	✓	✓	✓		$68.36_{\pm 0.31}$	$52.10_{\pm 0.36}$	$67.67_{\pm 0.17}$	$47.33_{\pm 0.65}$	$47.07_{\pm 0.36}$	$41.30_{\pm 0.24}$	$66.75_{\pm 0.92}$
(f)	✓	✓	✓	✓	✓	$\mathbf{69.01}_{\pm \mathbf{1.09}}$	$\mathbf{53.19}_{\pm \mathbf{0.38}}$	$\mathbf{68.43}_{\pm \mathbf{0.83}}$	$\mathbf{48.25}_{\pm \mathbf{0.59}}$	$\mathbf{47.93}_{\pm \mathbf{0.39}}$	$\mathbf{41.66}_{\pm \mathbf{0.15}}$	$\mathbf{66.80}_{\pm \mathbf{0.96}}$

表 9: 該表顯示了卷積融合器中每個部分的必要性。當所有部分都存在時，模型表現最佳。如果刪除其中任何一部分，模型表現都會變差。在該表中，「 V2TEx 」表示 V2T 提取器，「 T2VEn 」表示 T2V 編碼器，「 TrEn1 」表示 Transformer 編碼器 1，「 ConBl 」表示卷積塊，「 TrEn2 」表示 Transformer 編碼器 2。每列中的最佳結果以**粗體**突出顯示。

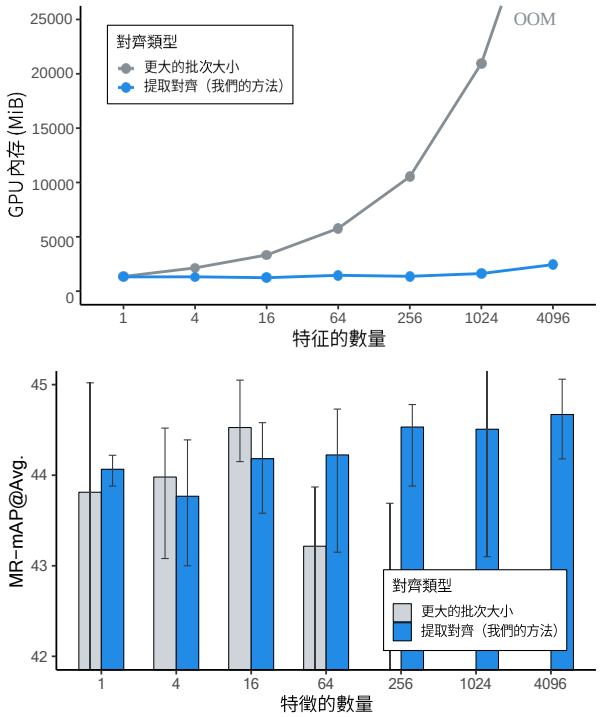

圖 7: 提取對齊方法可以在不佔用過多 GPU 內存的情況下引入更多樣本進行對比學習，隨著隊列長度的增加，模型效果會越來越好。我們可視化了兩種方法在對比學習所涉及的特徵數量增加時的 GPU 內存佔用情況以及在 QVHighlights 數據集上的結果。實驗在 NVIDIA GeForce RTX 4090 上進行，配備 2,4564MiB GPU 內存，以 TR-DETR (Sun et al. 2024) 基準模型。x 軸表示對比學習所涉及的特徵相對數量，其中 x 軸 = 1 表示批次大小爲 32，隊列長度爲 0。「 OOM 」表示內存不足。該圖展示的實驗與表 5 相同。

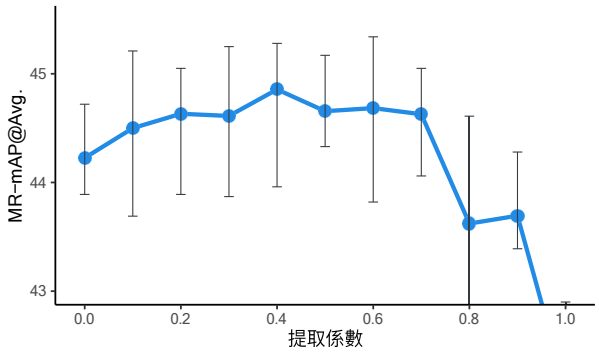

圖 8: 該圖顯示了提取對齊中提取係數對多模態比齊的影響。該圖顯示了與表 6 相同的一組實驗。

5 結論

在本文中，我們提出了一種用於視頻時刻檢索和高光檢測的模型 LD-DETR。我們首先提出了一種即插即用的方法提取對齊（Distill Align），它可以減輕重疊語義信息的影響。然後，我們引入了卷積融合器（Convolutional Fuser），它更能夠捕獲多模態特徵中的局部信息。最後，我們提出了一種即插即用的方法循環解碼器（Loop Decoder），它可以更充分地解碼多模態信息而不會導致過度擬合。我們的方法的優越性和有效性已在四個公共數據集上得到證明。

參考文獻

Anne Hendricks, L.; Wang, O.; Shechtman, E.; Sivic, J.; Darrell, T.; and Russell, B. 2017. Localizing moments in video with natural language. In Proceedings of the IEEE international conference on computer vision, 5803–5812.

Badamdorj, T.; Rochan, M.; Wang, Y.; and Cheng, L. 2021. Joint visual and audio learning for video highlight detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, 8127–8137.

Bahdanau, D.; Cho, K.; and Bengio, Y. 2014. Neu-

解碼器類型	解碼器大小	視頻時刻檢索					精彩片段檢測	
		R1		mAP			>=Very Good	
		@0.5	@0.7	@0.5	@0.75	Avg.	mAP	HIT@1
Moment-DETR (Lei, Berg, and Bansal 2021)								
更大的解碼器	1	$53.63_{\pm 1.59}$	$35.78_{\pm 0.25}$	$54.99_{\pm 1.33}$	$31.28_{\pm 1.07}$	$32.09_{\pm 1.01}$	$\mathbf{36.41}_{\pm \mathbf{0.40}}$	$\mathbf{56.92}_{\pm \mathbf{0.98}}$
	2	$54.71_{\pm 2.35}$	$35.84_{\pm 2.97}$	$55.80_{\pm 1.82}$	$31.39_{\pm 2.35}$	$32.50_{\pm 1.99}$	$36.17_{\pm 0.63}$	$56.63_{\pm 1.03}$
	3	$53.65_{\pm 0.73}$	$35.15_{\pm 1.28}$	$55.11_{\pm 0.68}$	$30.34_{\pm 0.63}$	$31.75_{\pm 0.73}$	$36.04_{\pm 0.37}$	$56.53_{\pm 0.98}$
	4	$52.72_{\pm 0.97}$	$35.14_{\pm 0.86}$	$54.65_{\pm 0.59}$	$30.64_{\pm 0.59}$	$31.59_{\pm 0.50}$	$35.78_{\pm 0.29}$	$55.48_{\pm 0.63}$
循環解碼器	1	$53.63_{\pm 1.59}$	$35.78_{\pm 0.25}$	$54.99_{\pm 1.33}$	$31.28_{\pm 1.07}$	$32.09_{\pm 1.01}$	$\mathbf{36.41}_{\pm \mathbf{0.40}}$	$\mathbf{56.92}_{\pm \mathbf{0.98}}$
	2	$55.69_{\pm 0.46}$	$37.63_{\pm 0.55}$	$\mathbf{56.92}_{\pm \mathbf{0.45}}$	$32.93_{\pm 0.36}$	$\mathbf{33.86}_{\pm \mathbf{0.16}}$	$36.37_{\pm 0.41}$	$56.88_{\pm 1.23}$
	3	$54.93_{\pm 1.14}$	$37.88_{\pm 1.28}$	$56.19_{\pm 0.76}$	$33.09_{\pm 0.81}$	$33.66_{\pm 0.45}$	$35.83_{\pm 0.16}$	$55.30_{\pm 0.60}$
	4	$\mathbf{55.70}_{\pm \mathbf{0.90}}$	$\mathbf{38.24}_{\pm \mathbf{0.85}}$	$56.74_{\pm 0.93}$	$\mathbf{33.19}_{\pm \mathbf{0.70}}$	$33.85_{\pm 0.66}$	$36.10_{\pm 0.43}$	$56.43_{\pm 0.54}$
BM-DETR (Jung et al. 2023)								
更大的解碼器	1	$60.90_{\pm 0.79}$	$44.10_{\pm 1.43}$	$60.91_{\pm 0.96}$	$39.03_{\pm 1.46}$	$38.93_{\pm 0.98}$	-	-
	2	$61.24_{\pm 1.33}$	$44.13_{\pm 1.42}$	$61.25_{\pm 0.88}$	$39.21_{\pm 1.06}$	$38.96_{\pm 1.03}$	-	-
	3	$60.72_{\pm 0.54}$	$44.14_{\pm 1.00}$	$61.14_{\pm 0.43}$	$39.74_{\pm 1.05}$	$39.18_{\pm 0.68}$	-	-
	4	$60.82_{\pm 0.89}$	$43.73_{\pm 0.78}$	$61.20_{\pm 0.84}$	$38.97_{\pm 0.80}$	$38.83_{\pm 0.81}$	-	-
循環解碼器	1	$60.90_{\pm 0.79}$	$44.10_{\pm 1.43}$	$60.91_{\pm 0.96}$	$39.03_{\pm 1.46}$	$38.93_{\pm 0.98}$	-	-
	2	$61.01_{\pm 0.70}$	$\mathbf{44.80}_{\pm \mathbf{0.96}}$	$61.68_{\pm 0.96}$	$\mathbf{40.34}_{\pm \mathbf{0.93}}$	$39.77_{\pm 0.75}$	-	-
	3	$61.42_{\pm 1.20}$	$44.70_{\pm 0.95}$	$61.71_{\pm 1.06}$	$39.75_{\pm 0.27}$	$39.24_{\pm 0.63}$	-	-
	4	$\mathbf{61.62}_{\pm \mathbf{0.73}}$	$44.76_{\pm 1.48}$	$\mathbf{61.74}_{\pm \mathbf{0.47}}$	$40.14_{\pm 0.92}$	$\mathbf{39.81}_{\pm \mathbf{0.87}}$	-	-
QD-DETR (Moon et al. 2023b)								
更大的解碼器	1	$61.98_{\pm 0.55}$	$47.30_{\pm 0.69}$	$62.03_{\pm 0.43}$	$41.96_{\pm 0.66}$	$41.42_{\pm 0.28}$	$38.92_{\pm 0.30}$	$62.05_{\pm 0.86}$
	2	$62.59_{\pm 0.70}$	$47.48_{\pm 0.79}$	$62.02_{\pm 0.75}$	$41.90_{\pm 0.63}$	$41.37_{\pm 0.54}$	$39.17_{\pm 0.21}$	$62.18_{\pm 0.67}$
	3	$61.73_{\pm 1.01}$	$45.38_{\pm 1.63}$	$60.45_{\pm 1.22}$	$39.99_{\pm 1.75}$	$39.57_{\pm 1.20}$	$\mathbf{39.24}_{\pm \mathbf{0.23}}$	$\mathbf{62.99}_{\pm \mathbf{0.74}}$
	4	$60.96_{\pm 0.43}$	$44.94_{\pm 1.03}$	$59.92_{\pm 0.40}$	$39.53_{\pm 0.76}$	$39.04_{\pm 0.77}$	$38.82_{\pm 0.27}$	$61.39_{\pm 0.93}$
循環解碼器	1	$61.98_{\pm 0.55}$	$47.30_{\pm 0.69}$	$62.03_{\pm 0.43}$	$41.96_{\pm 0.66}$	$41.42_{\pm 0.28}$	$38.92_{\pm 0.30}$	$62.05_{\pm 0.86}$
	2	$\mathbf{62.80}_{\pm \mathbf{0.80}}$	$47.82_{\pm 0.53}$	$62.97_{\pm 0.49}$	$42.93_{\pm 0.61}$	$42.26_{\pm 0.37}$	$39.23_{\pm 0.13}$	$62.83_{\pm 0.85}$
	3	$62.49_{\pm 0.37}$	$47.57_{\pm 0.44}$	$\mathbf{63.12}_{\pm \mathbf{0.47}}$	$42.68_{\pm 0.36}$	$42.32_{\pm 0.32}$	$38.94_{\pm 0.19}$	$61.88_{\pm 0.43}$
	4	$62.45_{\pm 0.86}$	$\mathbf{47.95}_{\pm \mathbf{0.42}}$	$63.03_{\pm 0.41}$	$\mathbf{43.20}_{\pm \mathbf{0.08}}$	$\mathbf{42.48}_{\pm \mathbf{0.24}}$	$39.14_{\pm 0.15}$	$62.67_{\pm 0.54}$
CG-DETR (Moon et al. 2023a)								
更大的解碼器	1	$65.92_{\pm 0.22}$	$50.44_{\pm 0.54}$	$65.50_{\pm 0.22}$	$45.62_{\pm 0.68}$	$44.76_{\pm 0.26}$	$40.34_{\pm 0.20}$	$65.12_{\pm 0.64}$
	2	$65.81_{\pm 1.23}$	$50.09_{\pm 0.54}$	$64.40_{\pm 0.31}$	$44.17_{\pm 0.51}$	$43.71_{\pm 0.37}$	$40.33_{\pm 0.25}$	$65.10_{\pm 0.40}$
	3	$65.30_{\pm 0.87}$	$49.22_{\pm 0.95}$	$63.50_{\pm 0.85}$	$43.66_{\pm 0.45}$	$42.85_{\pm 0.45}$	$\mathbf{40.45}_{\pm \mathbf{0.20}}$	$\mathbf{65.61}_{\pm \mathbf{0.50}}$
	4	$63.87_{\pm 2.51}$	$46.92_{\pm 4.35}$	$61.63_{\pm 3.07}$	$41.30_{\pm 3.54}$	$40.78_{\pm 3.12}$	$40.37_{\pm 0.11}$	$65.03_{\pm 0.57}$
循環解碼器	1	$65.92_{\pm 0.22}$	$50.44_{\pm 0.54}$	$65.50_{\pm 0.22}$	$45.62_{\pm 0.68}$	$44.76_{\pm 0.26}$	$40.34_{\pm 0.20}$	$65.12_{\pm 0.64}$
	2	$65.56_{\pm 0.49}$	$50.58_{\pm 0.59}$	$65.24_{\pm 0.44}$	$45.29_{\pm 0.59}$	$44.61_{\pm 0.39}$	$40.10_{\pm 0.28}$	$64.95_{\pm 0.79}$
	3	$\mathbf{66.26}_{\pm \mathbf{0.64}}$	$\mathbf{51.72}_{\pm \mathbf{0.53}}$	$\mathbf{65.74}_{\pm \mathbf{0.39}}$	$45.66_{\pm 0.27}$	$45.05_{\pm 0.32}$	$40.39_{\pm 0.24}$	$65.53_{\pm 0.47}$
	4	$66.22_{\pm 0.50}$	$51.04_{\pm 0.36}$	$66.01_{\pm 0.51}$	$\mathbf{45.95}_{\pm \mathbf{0.52}}$	$\mathbf{45.23}_{\pm \mathbf{0.35}}$	$40.33_{\pm 0.13}$	$65.50_{\pm 0.91}$
TR-DETR (Sun et al. 2024)								
更大的解碼器	1	$66.56_{\pm 1.06}$	$50.13_{\pm 0.89}$	$65.70_{\pm 0.79}$	$45.10_{\pm 0.78}$	$44.33_{\pm 0.51}$	$40.88_{\pm 0.19}$	$65.54_{\pm 0.45}$
	2	$66.13_{\pm 1.09}$	$49.54_{\pm 1.04}$	$64.02_{\pm 1.63}$	$43.86_{\pm 1.11}$	$43.35_{\pm 0.91}$	$40.61_{\pm 0.37}$	$65.10_{\pm 1.00}$
	3	$64.18_{\pm 1.22}$	$46.79_{\pm 1.25}$	$62.46_{\pm 0.96}$	$42.12_{\pm 0.91}$	$41.33_{\pm 0.85}$	$\mathbf{41.04}_{\pm \mathbf{0.16}}$	$65.28_{\pm 0.57}$
	4	$64.71_{\pm 0.94}$	$46.92_{\pm 1.35}$	$62.69_{\pm 0.71}$	$42.26_{\pm 1.25}$	$41.54_{\pm 0.86}$	$40.85_{\pm 0.10}$	$\mathbf{65.65}_{\pm \mathbf{0.88}}$
循環解碼器	1	$66.56_{\pm 1.06}$	$50.13_{\pm 0.89}$	$65.70_{\pm 0.79}$	$45.10_{\pm 0.78}$	$44.33_{\pm 0.51}$	$40.88_{\pm 0.19}$	$65.54_{\pm 0.45}$
	2	$66.59_{\pm 0.63}$	$51.10_{\pm 0.40}$	$65.97_{\pm 0.20}$	$45.42_{\pm 0.20}$	$44.90_{\pm 0.11}$	$40.93_{\pm 0.29}$	$65.60_{\pm 0.81}$
	3	$\mathbf{67.12}_{\pm \mathbf{0.36}}$	$51.08_{\pm 0.47}$	$\mathbf{66.35}_{\pm \mathbf{0.46}}$	$\mathbf{45.58}_{\pm \mathbf{0.47}}$	$\mathbf{44.92}_{\pm \mathbf{0.33}}$	$40.78_{\pm 0.44}$	$65.37_{\pm 0.80}$
	4	$66.94_{\pm 0.72}$	$\mathbf{51.40}_{\pm \mathbf{0.51}}$	$66.04_{\pm 0.69}$	$45.22_{\pm 0.64}$	$44.90_{\pm 0.34}$	$40.85_{\pm 0.11}$	$65.63_{\pm 0.69}$

表 10: 此頁空間太小，寫不開了。請參閱表格 11 瞭解詳情。

解碼器類型	解碼器大小	視頻時刻檢索					精彩片段檢測	
		R1		mAP			>=Very Good	
		@0.5	@0.7	@0.5	@0.75	Avg.	mAP	HIT@1
UVCOM (Xiao et al. 2024)								
更大的解碼器	1	$64.49_{\pm 1.10}$	$49.94_{\pm 0.72}$	$64.03_{\pm 1.01}$	$44.88_{\pm 0.53}$	$44.39_{\pm 0.61}$	$39.94_{\pm 0.18}$	$64.03_{\pm 0.25}$
	2	$64.37_{\pm 0.33}$	$48.73_{\pm 0.73}$	$62.85_{\pm 0.88}$	$43.52_{\pm 1.16}$	$42.73_{\pm 0.81}$	$39.97_{\pm 0.16}$	$63.42_{\pm 1.09}$
	3	$63.21_{\pm 1.36}$	$47.52_{\pm 1.81}$	$61.61_{\pm 1.19}$	$43.23_{\pm 1.35}$	$42.36_{\pm 1.24}$	$39.81_{\pm 0.13}$	$63.44_{\pm 0.15}$
	4	$62.68_{\pm 2.90}$	$46.17_{\pm 3.85}$	$60.74_{\pm 2.57}$	$41.40_{\pm 2.81}$	$40.99_{\pm 2.84}$	$39.90_{\pm 0.48}$	$63.72_{\pm 0.64}$
循環解碼器	1	$64.49_{\pm 1.10}$	$49.94_{\pm 0.72}$	$64.03_{\pm 1.01}$	$44.88_{\pm 0.53}$	$44.39_{\pm 0.61}$	$39.94_{\pm 0.18}$	$64.03_{\pm 0.25}$
	2	$63.81_{\pm 0.37}$	$49.50_{\pm 0.60}$	$63.60_{\pm 0.38}$	$44.44_{\pm 0.37}$	$44.03_{\pm 0.24}$	$39.81_{\pm 0.10}$	$63.37_{\pm 0.38}$
	3	$64.55_{\pm 0.69}$	$50.04_{\pm 0.95}$	$64.11_{\pm 0.54}$	$45.41_{\pm 1.90}$	$44.26_{\pm 0.56}$	$39.94_{\pm 0.17}$	$63.94_{\pm 0.46}$
	4	$\mathbf{64.65_{\pm 0.61}}$	$\mathbf{50.33_{\pm 0.55}}$	$\mathbf{64.17_{\pm 0.46}}$	$44.54_{\pm 0.44}$	$44.18_{\pm 0.29}$	$39.76_{\pm 0.18}$	$63.69_{\pm 0.80}$
循環解碼器 w/o SA	1	$64.36_{\pm 0.44}$	$50.21_{\pm 1.26}$	$63.99_{\pm 0.27}$	$\mathbf{45.52_{\pm 0.68}}$	$\mathbf{44.72_{\pm 0.53}}$	$39.85_{\pm 0.21}$	$63.82_{\pm 1.05}$
	2	$64.36_{\pm 0.54}$	$50.06_{\pm 0.80}$	$63.85_{\pm 0.30}$	$45.15_{\pm 0.68}$	$44.43_{\pm 0.51}$	$39.88_{\pm 0.24}$	$\mathbf{64.19_{\pm 0.59}}$
	3	$64.00_{\pm 0.64}$	$49.30_{\pm 1.22}$	$63.73_{\pm 0.54}$	$44.33_{\pm 1.00}$	$43.93_{\pm 0.63}$	$39.79_{\pm 0.20}$	$63.54_{\pm 0.74}$
	4	$63.39_{\pm 0.61}$	$49.02_{\pm 0.50}$	$63.32_{\pm 0.31}$	$43.88_{\pm 0.52}$	$43.20_{\pm 0.35}$	$39.35_{\pm 0.18}$	$62.59_{\pm 0.09}$
循環解碼器 w/o CTA	1	$63.96_{\pm 0.37}$	$49.73_{\pm 0.42}$	$63.65_{\pm 0.40}$	$44.55_{\pm 0.47}$	$44.12_{\pm 0.24}$	$39.82_{\pm 0.20}$	$63.63_{\pm 0.94}$
	2	$\mathbf{64.65_{\pm 0.53}}$	$50.09_{\pm 0.59}$	$\mathbf{64.17_{\pm 0.72}}$	$44.79_{\pm 0.87}$	$44.42_{\pm 0.61}$	$39.83_{\pm 0.17}$	$63.90_{\pm 0.41}$
	3	$64.62_{\pm 0.77}$	$50.10_{\pm 0.40}$	$64.12_{\pm 0.64}$	$44.74_{\pm 0.44}$	$44.25_{\pm 0.35}$	$\mathbf{39.98_{\pm 0.15}}$	$63.74_{\pm 0.36}$
	4	$63.85_{\pm 0.78}$	$49.55_{\pm 0.45}$	$63.60_{\pm 1.13}$	$44.87_{\pm 0.67}$	$44.15_{\pm 0.82}$	$35.72_{\pm 7.91}$	$63.54_{\pm 0.67}$

表 11: 循環解碼器作爲一種即插即用的方法，可以幫助多個模型獲得更好的結果，而不會像更大的解碼器那樣出現過擬合的風險。該表按基準模型對模型進行分類。對於循環解碼器，解碼器大小表示循環解碼器循環的數量，對於更大的解碼器，解碼器大小表示解碼器層數的倍數。插槽注意力（Slot Attention，SA）和片段文本對齊（Clip-Text Alignment，CTA）是 UVCOM (Xiao et al. 2024) 模型中使用的兩種方法。每列中每類特徵中的最佳結果以**粗體**突出顯示。該表顯示了與圖 9 相同的一組實驗。

ral machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473.

Cai, S.; Zuo, W.; Davis, L. S.; and Zhang, L. 2018. Weakly-supervised video summarization using variational encoder-decoder and web prior. In Proceedings of the European conference on computer vision (ECCV), 184–200.

Chen, G.; Huang, Y.; Xu, J.; Pei, B.; Chen, Z.; Li, Z.; Wang, J.; Li, K.; Lu, T.; and Wang, L. 2024. Video mamba suite: State space model as a versatile alternative for video understanding. arXiv preprint arXiv:2403.09626.

Chen, J.; Chen, X.; Ma, L.; Jie, Z.; and Chua, T.-S. 2018. Temporally grounding natural sentence in video. In Proceedings of the 2018 conference on empirical methods in natural language processing, 162–171.

Chen, S.; and Jiang, Y.-G. 2019. Semantic proposal for activity localization in videos via sentence query. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 33, 8199–8206.

Chung, J.; Gulcehre, C.; Cho, K.; and Bengio, Y. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555.

Escorcia, V.; Soldan, M.; Sivic, J.; Ghanem, B.; and Russell, B. 2019a. Finding moments in video collections using natural language. arXiv preprint arXiv:1907.12763.

Escorcia, V.; Soldan, M.; Sivic, J.; Ghanem, B.; and Russell, B. 2019b. Temporal localization of moments in video collections with natural language.

Feichtenhofer, C.; Fan, H.; Malik, J.; and He, K. 2019. Slowfast networks for video recognition. In Proceedings of the IEEE/CVF international conference on computer vision, 6202–6211.

Gao, J.; Sun, C.; Yang, Z.; and Nevatia, R. 2017. Tall: Temporal activity localization via language query. In Proceedings of the IEEE international conference on computer vision, 5267–5275.

Gao, J.; Sun, X.; Xu, M.; Zhou, X.; and Ghanem, B. 2021a. Relation-aware video reading comprehension for temporal language grounding. arXiv preprint arXiv:2110.05717.

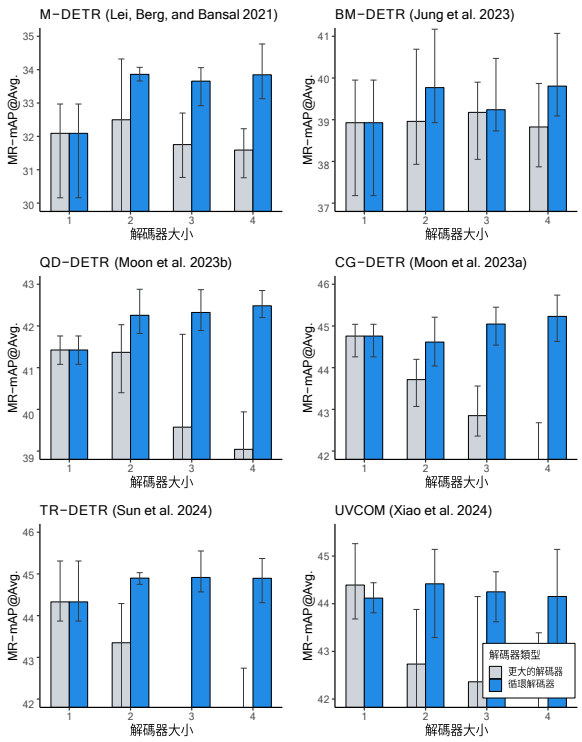

圖 9: 循環解碼器作爲一種即插即用的方法，可以幫助多個模型實現更好的結果，而不會出現過擬合的風險。對於循環解碼器，x 軸表示循環解碼器循環的數量。對於更大的解碼器解碼器，x 軸表示解碼器層數的倍數。此圖顯示了與表 10 和表 11 相同的一組實驗。

Gao, P.; Zheng, M.; Wang, X.; Dai, J.; and Li, H. 2021b. Fast convergence of detr with spatially modulated co-attention. In Proceedings of the IEEE/CVF international conference on computer vision, 3621–3630.

Gygli, M.; Song, Y.; and Cao, L. 2016. Video2gif: Automatic generation of animated gifs from video. In Proceedings of the IEEE conference on computer vision and pattern recognition, 1001–1009.

Han, D.; Cheng, X.; Guo, N.; Ye, X.; Rainer, B.; and Priller, P. 2023. Momentum cross-modal contrastive learning for video moment retrieval. IEEE Transactions on Circuits and Systems for Video Technology.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 9729–9738.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, 770–778.

Hoffmann, D. T.; Behrmann, N.; Gall, J.; Brox, T.; and Noroozi, M. 2022. Ranking info noise contrastive estimation: Boosting contrastive learning via ranked positives. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 36, 897–905.

Hong, F.-T.; Huang, X.; Li, W.-H.; and Zheng, W.-S. 2020. Mini-net: Multiple instance ranking network for video highlight detection. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16, 345–360. Springer.

Huang, C.; Wu, Y.-L.; Shuai, H.-H.; and Huang, C.-C. 2024. Semantic Fusion Augmentation and Semantic Boundary Detection: A Novel Approach to Multi-Target Video Moment Retrieval. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 6783–6792.

Jang, J.; Park, J.; Kim, J.; Kwon, H.; and Sohn, K. 2023. Knowing where to focus: Event-aware transformer for video grounding. In Proceedings of the IEEE/CVF International Conference on Computer Vision, 13846–13856.

Ji, W.; Liang, R.; Zheng, Z.; Zhang, W.; Zhang, S.; Li, J.; Li, M.; and Chua, T.-s. 2023. Are binary annotations sufficient? video moment retrieval via hierarchical uncertainty-based active learning. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 23013–23022.

Jiang, Y.; Zhang, W.; Zhang, X.; Wei, X.; Chen, C. W.; and Li, Q. 2024. Prior Knowledge Integration via LLM Encoding and Pseudo Event Regulation for Video Moment Retrieval. arXiv preprint arXiv:2407.15051.

Jing, W.; Sun, A.; Zhang, H.; and Li, X. 2023. MS-DETR: Natural Language Video Localization with Sampling Moment-Moment Interaction. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 1387–1400. Toronto, Canada: Association for Computational Linguistics.

Jung, M.; Jang, Y.; Choi, S.; Kim, J.; Kim, J.-H.; and Zhang, B.-T. 2023. Overcoming Weak Visual-Textual Alignment for Video Moment Retrieval. arXiv preprint arXiv:2306.02728.

Kong, Q.; Cao, Y.; Iqbal, T.; Wang, Y.; Wang, W.; and Plumbley, M. D. 2020. Panns: Large-scale pretrained audio neural networks for audio pattern recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 28: 2880–2894.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, 25.

LeCun, Y.; Bottou, L.; Bengio, Y.; and Haffner, P. 1998. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11): 2278–2324.

Lee, P.; and Byun, H. 2023. BAM-DETR: Boundary-Aligned Moment Detection Transformer for Temporal Sentence Grounding in Videos. arXiv preprint arXiv:2312.00083.

Lei, J.; Berg, T. L.; and Bansal, M. 2021. Detecting moments and highlights in videos via natural language queries. Advances in Neural Information Processing Systems, 34: 11846–11858.

Lei, J.; Yu, L.; Berg, T. L.; and Bansal, M. 2020. Tvr: A large-scale dataset for video-subtitle moment retrieval. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16, 447–463. Springer.

Li, J.; Zhang, F.; Lin, S.; Zhou, F.; and Wang, R. 2023. Mim: Lightweight multi-modal interaction model for joint video moment retrieval and highlight detection. In 2023 IEEE International Conference on Multimedia and Expo (ICME), 1961–1966. IEEE.

Li, K.; Guo, D.; and Wang, M. 2021. Proposal-free video grounding with contextual pyramid network. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 35, 1902–1910.

Li, P.; Xie, C.-W.; Xie, H.; Zhao, L.; Zhang, L.; Zheng, Y.; Zhao, D.; and Zhang, Y. 2024. Momentdiff: Generative video moment retrieval from random to real. Advances in neural information processing systems, 36.

Lin, K. Q.; Zhang, P.; Chen, J.; Pramanick, S.; Gao, D.; Wang, A. J.; Yan, R.; and Shou, M. Z. 2023. Univtg: Towards unified video-language temporal grounding. In Proceedings of the IEEE/CVF International Conference on Computer Vision, 2794–2804.

Liu, D.; Qu, X.; Dong, J.; Zhou, P.; Cheng, Y.; Wei, W.; Xu, Z.; and Xie, Y. 2021. Context-aware biaffine localizing network for temporal sentence grounding. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 11235–11244.

Liu, M.; Wang, X.; Nie, L.; Tian, Q.; Chen, B.; and Chua, T.-S. 2018. Cross-modal moment localization in videos. In Proceedings of the 26th ACM international conference on Multimedia, 843–851.

Liu, S.; Li, F.; Zhang, H.; Yang, X.; Qi, X.; Su, H.; Zhu, J.; and Zhang, L. 2022a. Dab-detr: Dynamic anchor boxes are better queries for detr. arXiv preprint arXiv:2201.12329.

Liu, W.; Mei, T.; Zhang, Y.; Che, C.; and Luo, J. 2015. Multi-task deep visual-semantic embedding for video thumbnail selection. In Proceedings of the IEEE conference on computer vision and pattern recognition, 3707–3715.

Liu, W.; Miao, B.; Cao, J.; Zhu, X.; Liu, B.; Nasim, M.; and Mian, A. 2024a. Context-Enhanced Video Moment Retrieval with Large Language Models. arXiv preprint arXiv:2405.12540.

Liu, Y.; He, J.; Li, W.; Kim, J.; Wei, D.; Pfister, H.; and Chen, C. W. 2024b. *R*ˆ2-Tuning: Efficient Image-to-Video Transfer Learning for Video Temporal Grounding. arXiv preprint arXiv:2404.00801.

Liu, Y.; Li, S.; Wu, Y.; Chen, C.-W.; Shan, Y.; and Qie, X. 2022b. Umt: Unified multi-modal transformers for joint video moment retrieval and highlight detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 3042–3051.

Liu, Z.; Li, J.; Xie, H.; Li, P.; Ge, J.; Liu, S.-A.; and Jin, G. 2024c. Towards balanced alignment: Modal-enhanced semantic modeling for video moment retrieval. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 38, 3855–3863.

Loshchilov, I. 2017. Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101.

Ma, K.; Fang, H.; Zang, X.; Ban, C.; Zhou, L.; He, Z.; Li, Y.; Sun, H.; Feng, Z.; and Hou, X. 2024. Disentangle and denoise: Tackling context misalignment for video moment retrieval. arXiv preprint arXiv:2408.07600.

Ma, K.; Zang, X.; Feng, Z.; Fang, H.; Ban, C.; Wei, Y.; He, Z.; Li, Y.; and Sun, H. 2023. LLaViLo: Boosting Video Moment Retrieval via Adapter-Based Multimodal Modeling. In Proceedings of the IEEE/CVF International Conference on Computer Vision, 2798–2803.

Ma, Y.; Yang, T.; Shan, Y.; and Li, X. 2022. Simvtp: Simple video text pre-training with masked autoencoders. arXiv preprint arXiv:2212.03490.

Meng, D.; Chen, X.; Fan, Z.; Zeng, G.; Li, H.; Yuan, Y.; Sun, L.; and Wang, J. 2021. Conditional detr for fast training convergence. In Proceedings of the IEEE/CVF international conference on computer vision, 3651–3660.

Moon, W.; Hyun, S.; Lee, S.; and Heo, J.-P. 2023a. Correlation-guided Query-Dependency Calibration in Video Representation Learning for Temporal Grounding. arXiv preprint arXiv:2311.08835.

Moon, W.; Hyun, S.; Park, S.; Park, D.; and Heo, J.-P. 2023b. Query-dependent video representation for moment retrieval and highlight detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 23023–23033.

Nan, G.; Qiao, R.; Xiao, Y.; Liu, J.; Leng, S.; Zhang, H.; and Lu, W. 2021. Interventional video grounding with dual contrastive learning. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2765–2775.

Panta, L.; Shrestha, P.; Sapkota, B.; Bhattarai, A.; Manandhar, S.; and Sah, A. K. 2024. Cross-modal Contrastive Learning with Asymmetric Co-attention Network for Video Moment Retrieval. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 607–614.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In International conference on machine learning, 8748–8763. PMLR.

Regneri, M.; Rohrbach, M.; Wetzel, D.; Thater, S.; Schiele, B.; and Pinkal, M. 2013. Grounding action descriptions in videos. Transactions of the Association for Computational Linguistics, 1: 25–36.

Rezatofighi, H.; Tsoi, N.; Gwak, J.; Sadeghian, A.; Reid, I.; and Savarese, S. 2019. Generalized intersection over union: A metric and a loss for bounding box regression. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 658–666.

Rumelhart, D. E.; Hinton, G. E.; and Williams, R. J. 1986. Learning representations by back-propagating errors. nature, 323(6088): 533–536.

Soldan, M.; Xu, M.; Qu, S.; Tegner, J.; and Ghanem, B. 2021. Vlg-net: Video-language graph matching network for video grounding. In Proceedings of the IEEE/CVF International Conference on Computer Vision, 3224–3234.

Song, Y.; Redi, M.; Vallmitjana, J.; and Jaimes, A. 2016. To click or not to click: Automatic selection of beautiful thumbnails from videos. In Proceedings of the 25th ACM international on conference on information and knowledge management, 659–668.

Song, Y.; Vallmitjana, J.; Stent, A.; and Jaimes, A. 2015. Tvsum: Summarizing web videos using titles. In Proceedings of the IEEE conference on computer vision and pattern recognition, 5179–5187.

Sun, H.; Zhou, M.; Chen, W.; and Xie, W. 2024. TR-DETR: Task-Reciprocal Transformer for Joint Moment Retrieval and Highlight Detection. arXiv preprint arXiv:2401.02309.

Tian, Y.; Krishnan, D.; and Isola, P. 2020. Contrastive multiview coding. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16, 776–794. Springer.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, F.; and Polosukhin, I. 2017. Attention is all you need. Advances in neural information processing systems, 30.

Wang, L.; Liu, D.; Puri, R.; and Metaxas, D. N. 2020. Learning trailer moments in full-length movies with co-contrastive attention. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16, 300–316. Springer.

Wang, R.; Feng, J.; Zhang, F.; Luo, X.; and Luo, Y. 2024. Modality-aware Heterogeneous Graph for Joint Video Moment Retrieval and Highlight Detection. IEEE Transactions on Circuits and Systems for Video Technology.

Wang, T.; Zhang, J.; Zheng, F.; Jiang, W.; Cheng, R.; and Luo, P. 2023. Learning grounded vision-language representation for versatile understanding in untrimmed videos. arXiv preprint arXiv:2303.06378.

Wang, W.; Huang, Y.; and Wang, L. 2019. Language-driven temporal activity localization: A semantic matching reinforcement learning model. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 334–343.

Wang, Y.; Zhang, X.; Yang, T.; and Sun, J. 2022a. Anchor detr: Query design for transformer-based detector. In Proceedings of the AAAI conference on artificial intelligence, volume 36, 2567–2575.

Wang, Z.; Wang, L.; Wu, T.; Li, T.; and Wu, G. 2022b. Negative sample matters: A renaissance of metric learning for temporal grounding. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 36, 2613–2623.

Wei, F.; Wang, B.; Ge, T.; Jiang, Y.; Li, W.; and Duan, L. 2022. Learning pixel-level distinctions for video highlight detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 3073–3082.

Xiao, S.; Chen, L.; Zhang, S.; Ji, W.; Shao, J.; Ye, L.; and Xiao, J. 2021. Boundary proposal network for two-stage natural language video localization. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 35, 2986–2994.

Xiao, Y.; Luo, Z.; Liu, Y.; Ma, Y.; Bian, H.; Ji, Y.; Yang, Y.; and Li, X. 2024. Bridging the gap: A unified video comprehension framework for moment retrieval and highlight detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 18709–18719.

Xiong, B.; Kalantidis, Y.; Ghadiyaram, D.; and Grauman, K. 2019. Less is more: Learning highlight detection from video duration. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 1258–1267.

Xu, M.; Wang, H.; Ni, B.; Zhu, R.; Sun, Z.; and Wang, C. 2021. Cross-category video highlight detection via set-based learning. In Proceedings of the IEEE/CVF International Conference on Computer Vision, 7970–7979.

Xu, Y.; Sun, Y.; Zhai, B.; Jia, Y.; and Du, S. 2024. Mh-detr: Video moment and highlight detection with cross-modal transformer. In 2024 International Joint Conference on Neural Networks (IJCNN), 1–8. IEEE.

Yan, S.; Xiong, X.; Nagrani, A.; Arnab, A.; Wang, Z.; Ge, W.; Ross, D.; and Schmid, C. 2023. Unloc: A unified framework for video localization tasks. In Proceedings of the IEEE/CVF International Conference on Computer Vision, 13623–13633.

Yang, J.; Wei, P.; Li, H.; and Ren, Z. 2024. Task-Driven Exploration: Decoupling and Inter-Task Feedback for Joint Moment Retrieval and Highlight Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 18308–18318.

Yao, T.; Mei, T.; and Rui, Y. 2016. Highlight detection with pairwise deep ranking for first-person video summarization. In Proceedings of the IEEE conference on computer vision and pattern recognition, 982–990.

Yao, Z.; Ai, J.; Li, B.; and Zhang, C. 2021. Efficient detr: improving end-to-end object detector with dense prior. arXiv preprint arXiv:2104.01318.

Ye, Q.; Shen, X.; Gao, Y.; Wang, Z.; Bi, Q.; Li, P.; and Yang, G. 2021. Temporal cue guided video highlight detection with low-rank audio-visual fusion. In Proceedings of the IEEE/CVF International Conference on Computer Vision, 7950–7959.

Yu, S.; Cho, J.; Yadav, P.; and Bansal, M. 2024. Self-chained image-language model for video localization and question answering. Advances in Neural Information Processing Systems, 36.

Yu, Y.; Lee, S.; Na, J.; Kang, J.; and Kim, G. 2018. A deep ranking model for spatio-temporal highlight detection from a 360F video. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 32.

Yuan, L.; Tay, F. E. H.; Li, P.; and Feng, J. 2019a. Unsupervised video summarization with cycle-consistent adversarial LSTM networks. IEEE Transactions on Multimedia, 22(10): 2711–2722.

Yuan, Y.; Ma, L.; Wang, J.; Liu, W.; and Zhu, W. 2019b. Semantic conditioned dynamic modulation for temporal sentence grounding in videos. Advances in Neural Information Processing Systems, 32.

Yuan, Y.; Mei, T.; and Zhu, W. 2019. To find where you talk: Temporal sentence localization in video with attention based location regression. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 33, 9159–9166.

Zeng, R.; Xu, H.; Huang, W.; Chen, P.; Tan, M.; and Gan, C. 2020. Dense regression network for video grounding. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 10287–10296.

Zhang, D.; Dai, X.; Wang, X.; Wang, Y.-F.; and Davis, L. S. 2019. Man: Moment alignment network for natural

language moment retrieval via iterative graph adjustment. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 1247–1257.

Zhang, H.; Sun, A.; Jing, W.; and Zhou, J. T. 2020a. Span-based localizing network for natural language video localization. arXiv preprint arXiv:2004.13931.

Zhang, K.; Chao, W.-L.; Sha, F.; and Grauman, K. 2016. Video summarization with long short-term memory. In Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VII 14, 766–782. Springer.

Zhang, M.; Yang, Y.; Chen, X.; Ji, Y.; Xu, X.; Li, J.; and Shen, H. T. 2021. Multi-stage aggregated transformer network for temporal language localization in videos. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 12669–12678.

Zhang, S.; Peng, H.; Fu, J.; and Luo, J. 2020b. Learning 2d temporal adjacent networks for moment localization with natural language. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, 12870–12877.

Zheng, M.; Gao, P.; Zhang, R.; Li, K.; Wang, X.; Li, H.; and Dong, H. 2020. End-to-end object detection with adaptive clustering transformer. arXiv preprint arXiv:2011.09315.

Zhou, S.; Zhang, F.; Wang, R.; Zhou, F.; and Su, Z. 2024. Subtask Prior-driven Optimized Mechanism on Joint Video Moment Retrieval and Highlight Detection. IEEE Transactions on Circuits and Systems for Video Technology.

Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; and Dai, J. 2020. Deformable detr: Deformable transformers for end-to-end object detection. arXiv preprint arXiv:2010.04159.

楊金福; 劉玉斌; 宋琳; and 閆雪. 2022. 基於顯著特徵增強的跨模態視頻片段檢索. 電子與信息學報, 44(12): 4395–4404.

繆翌; 張衛鋒; and 徐領. 2024. 基於 CLIP 的視頻時刻檢索預訓練模型. 計算機應用研究, 1–8.

蔣尋; 徐行; 沈復民; 王國慶; and 楊陽. 2023. 無模態融合的高效弱監督視頻時刻檢索算法. 北京航空航天大學學報, 1–12.

陳卓; 杜昊; 吳雨菲; 徐童; and 陳恩紅. 2020. 基於視覺–文字關係對齊的跨模態視頻片段檢索. 中國科學: 信息科學, 50(06): 862–876.

黎金宇. 2024. 基於跨模態信息交互的視頻內容定位研究. 碩士論文, 中山大學.